Applied General Statistics

When you cannot measure what you are speaking about, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely in your thoughts advanced to the stage of a *science*, whatever the matter may be.

—*Lord Kelvin.*

*   *   *

When the facts are gathered or discovered, when they are disentangled and identified, when they are sifted and verified, when they are counted and measured, the real task of the scholar is not ended —it is not even begun, but only prepared for him.

—*MacIver.*

Applied
General Statistics

By

Frederick E. Croxton, Ph. D

Professor of Statistics, Columbia University

and

Dudley J. Cowden, Ph. D

Professor of Economic Statistics
School of Commerce
University of North Carolina

New York
Prentice-Hall, Inc.

PREFACE

This book is intended for the use of readers who are interested in the understanding of statistical methods, and in their application in various fields, especially the social sciences. Consequently, the illustrative material has been drawn mainly from the fields of economics, sociology, and business, but occasionally from others.

The arrangement of the topics treated in this volume is about the same as in the authors' earlier book, *Practical Business Statistics*. The present text, however, does not stress business applications of statistical methods, but does present a greatly amplified treatment of analytical methods. The extensive discussion of the description and analysis of statistical data and of the making of statistical inferences will, we hope, make it useful to a wide group of teachers emphasizing various aspects of statistics.

Although the treatment of the material in this text is intended to be at an elementary level, probably its scope is so great that it cannot be covered adequately in most introductory courses. Thus, many instructors will deem it advisable to omit certain material in the first course and reserve it for a second, more advanced or more specialized course. Among the chapters which may be completely omitted in a short course without disturbing the continuity of treatment are Chapters XIII, XVI, XVIII, XXIII, and XXIV. Parts of other chapters may, of course, be left out at the discretion of the instructor. In the table of contents, chapters or sections which could well be omitted from an elementary course have been starred.

One problem of the statistician, especially the teacher, is that of selecting symbols which are simple and clear. Some of the symbols used in the present volume differ from those used in the authors'

earlier text.   This departure was made in an attempt to arrive at symbols which would be more easily understood and which would, consequently, facilitate the teaching process.

If this volume has any merit, it must be at least partly because of those who first introduced the authors to statistical methods; because of those texts (and other publications) which have preceded this; because of those publishers and individuals who have allowed us to reproduce charts or data of particular value, specific acknowledgment of which is made in the appropriate connection; and because of numerous individuals who have assisted with the task of completing the book.   The authors especially extend their thanks to Mr. Morton S. Nagelberg, who assisted in the construction of charts and in connection with certain of the mathematical developments; to Dr. James D. Paris, who assisted and advised concerning a number of charts; to Brant Bonner, Herbert Wolf, and John W. Gunter, for aid in making computations and preparing charts; and to Rosetta R. Croxton, for assistance in the reading of proof.

<div align="right">

FREDERICK E. CROXTON
DUDLEY J. COWDEN

</div>

CONTENTS

(For a short course in statistical methods, the chapters or sections which are starred in this table of contents may be omitted without destroying the continuity of treatment.)

Applied General Statistics

CHAPTER I

INTRODUCTION

Statistical Data and Statistical Methods

The term *statistics* is used in either of two senses. In common parlance it is generally used synonymously with the term *data*. Thus someone may say that he has seen "statistics of industrial accidents in the United States." It would be conducive to greater precision of meaning if we were not to use statistics in this sense, but rather to say "data of (or figures concerning) industrial accidents in the United States."

"Statistics" also refers to the statistical principles and methods which have been developed for handling numerical data and which form the subject matter of this text. Statistical methods (or statistics) range from the most elementary descriptive devices, which may be understood by anyone, to those extremely complicated mathematical procedures which are comprehended by only the most expert theoreticians. It is the purpose of this volume not to enter into the highly mathematical and theoretical aspects of the subject but rather to treat of its more elementary and more frequently used phases.

Statistics (that is, statistical methods) may be defined as *the collection, presentation, analysis, and interpretation of numerical data*. The facts which are dealt with must be capable of numerical expression. We can make little use statistically of the information that dwellings are built of brick, stone, wood, etc.; however, if we are able to determine *how many* or *what proportion* of, dwellings are constructed of each type of material, we have numerical data useful for statistical analysis.

Statistics should not be thought of as a subject correlative with physics, chemistry, economics, and sociology. Statistics is not a science; it is a scientific method. The methods and procedures which we are about to examine constitute a useful and often indispensable tool for the research worker. Without an adequate understanding of statistics the investigator in the social sciences may frequently be like a blind man groping in a dark closet for a black cat that isn't there. The methods of statistics are useful

1

in an ever-widening range of human activities, in any field of thought in which numerical data may be had.

The derivation of the word "statistics" suggests its origin. The administration of states required the collection and analysis of data of population and wealth for purposes of war and finance. Gradually data of more diverse nature were obtained for the general uses of government. Certain phases of statistics were developed by students of games of chance. Insurance and biology, as well as other natural sciences, were fertile fields for the application and development of statistical methods. Today there is hardly a phase of human activity which does not find statistical devices at least occasionally useful. Economics, sociology, anthropology, business, agriculture, psychology, and education—all lean heavily upon statistics. The medical research worker often must rely upon statistics to determine the significance of his results. The lawyer, especially if he be in corporation practice, may frequently find statistical devices of definite use. It should, of course, be added that the musician, the artist, the actor, and the writer of fiction would rarely have occasion to use statistics, but even here certain data of sales, box-office receipts, and trends of popular taste might be apropos.

In defining statistics it was pointed out that the numerical data are collected, presented, analyzed, and interpreted. Let us briefly examine each of these four procedures.

Collection. Statistical data may be obtained from existing published or unpublished sources, such as government agencies, trade associations, research bureaus, magazines, newspapers, individual researchers, and elsewhere. On the other hand, the investigator may collect his own information, going perhaps from house to house to obtain the data. The first-hand collection of statistical data is one of the most difficult and important tasks which a statistician must face. The soundness of his procedure determines in an overwhelming degree the usefulness of the data which he obtains.

The following chapter treats of these two methods of obtaining data. It should be emphasized, however, that the investigator who has experience and good common sense is at a distinct advantage if original data must be collected. There is much which may be taught about this phase of statistics, but there is much more which can be learned only through experience. Although a person may never collect statistical data for his own use and may always use published sources, it is essential that he have a working knowledge of the processes of collection and that he be able to evaluate the reliability of the data he proposes to use. Unreliable data do not constitute a satisfactory base upon which to rest a conclusion.

It is to be regretted that many people have a tendency to accept sta-

tistical data without question. To them, any information which is presented statistically is regarded as correct; the mere fact that a statement has been put in definite quantitative terms is sufficient to establish its authenticity. For this reason it behooves the research worker to do his work with the greatest possible care, in order that his conclusions may be valid. Above all, the basic data which are collected must be as accurate and comprehensive as is feasible. Let it not be said, as Stuart Chase commented in a book review:

> The learned economists today make graphs, charts, index numbers, of things which they have inadequately observed. Their mathematics has run ahead of their science.

Presentation. Either for one's own use or for the use of others, the data must be presented in some suitable form. Usually the figures are arranged in tables or represented by graphic devices as described in Chapters III to VI.

Analysis. In the process of analysis, data must be classified into useful and logical categories. The possible categories must be considered when plans are made for collecting the data, and the data must be classified as they are tabulated and before they can be shown graphically. Thus the process of analysis is partially concurrent with collection and presentation.

There are four important bases of classification of statistical data: (1) qualitative, (2) quantitative, (3) chronological, and (4) geographical, each of which will be examined in turn.

Qualitative. When, for example, employees are classified as union or non-union, we have a qualitative differentiation. The distinction is one of kind rather than of amount. Individuals may be classified concerning marital status, as single, married, widowed, divorced, and separated. Farm operators may be classified as full owners, part owners, managers, and tenants. Rubber may be designated as plantation or wild, according to its source.

Quantitative. When items vary in respect to some measurable characteristic, a quantitative classification is appropriate. The United States Census of 1930 reports 12,351,549 non-farm homes which were rented. The distribution of monthly rentals is shown in Table 1.

Families may be classified according to the number of children. Manufacturing concerns may be classified according to the number of workers employed, and also according to the value of goods produced.

Most quantitative distributions are *frequency distributions.* The data of rented non-farm homes show the number (frequency) of homes falling in each rental category. Similarly, the data of Table 27 show a frequency distribution of the grades of the 1937 class of the United States Naval

TABLE 1

NON-FARM HOMES IN THE UNITED STATES
CLASSIFIED BY AMOUNT OF MONTHLY
RENTAL, 1930

Monthly rental	Number of non-farm homes
Under $10	1,563,952
$10.00 to 14 99	1,330,927
15 00 to 19 99	1,302,387
20 00 to 29 99	2,545,208
30 00 to 49.99	3,191,435
50 00 to 74.99	1,503,401
75.00 to 99.99	343,071
100.00 and over	255,339
Rental unknown	315,829
Total	12,351,549

Source *Statistical Abstract of the United States, 1937*, p. 50 and by correspondence

Academy. A number of other frequency distributions are shown in Chapters VIII, IX, and X.

Sometimes, qualitatively classified data may be reclassified on a quantitative basis by making very slight changes. The assets of a bank may be listed in respect to degree of liquidity (cash, due from banks, United States securities, marketable securities, call loans, eligible paper, other loans, real estate loans, real estate, and furniture and fixtures). Although these categories differ from one another in a more or less unassignable quantitative fashion, the classification is actually made upon a qualitative basis. If we should reclassify the bank assets according to the length of time required to convert each into cash, the classification would be quantitative. In general the assets would be in the same order as before, but a few specific items among the less liquid qualitative groups (for example, certain real estate and real estate loans) would be convertible into cash in a relatively short time.

Chronological. Chronological data or *time series* show figures concerning a particular phenomenon at various specified times. For example, the closing price of a certain stock may be shown for each day over a period of months or years; the birth rate in the United States may be listed for each of a number of years; production of coal may be shown monthly for a span of years. The analysis of time series, involving a consideration of trend, cyclical, periodic (seasonal), and accidental movements, will be discussed in Chapters XIV to XIX.

In a certain sense, time series are somewhat akin to quantitative distri-

butions in that each succeeding year or month of a series is one year further removed from some earlier point of reference. However, periods of time— or, rather, the events occurring within these periods—differ qualitatively from each other. The essential arrangement of the figures in a time se-quence is inherent in the nature of the data under consideration.

Occasionally a time series may be converted into a frequency distribu-tion. If a railroad company has kept records of the number of railroad ties replaced each year, the data constitute a time series. When the same information is used in conjunction with the dates of installation, the life of the various ties may be expressed as a frequency distribution, showing perhaps:

Length of life	Number of ties
4 but under 5 years	2
5 but under 6 years	5
6 but under 7 years.... .	17
etc.	etc.

Geographical. The geographical distribution is essentially a type of qualitative distribution, but is generally considered as a distinct classifica-tion. When the population is shown for each of the states in the United States, we have data which are classified geographically. Although there is a qualitative difference between any two states, the distinction that is being made is not one of kind but of location. Various geographical series are shown in Tables 4 and 7 and in Chart 57.

Sometimes a geographical distribution may be put into the form of a frequency distribution. Thus, if we had data of the yield of corn per acre in each county of Iowa, we should have a geographical series. This may be put into the form of a frequency distribution by stating the number of counties having yields per acre of "10 and under 15 bushels," "15 and under 20 bushels," etc.

The presentation of classified data in tabular and graphic form is but one elementary step in the analysis of statistical data. Many other proc-esses are described in the following pages of this book. Statistical investi-gation frequently endeavors to ascertain what is typical in a given situa-tion. Hence all types of occurrences must be considered, both the usual and the unusual.

In forming an opinion, most individuals are apt to be unduly influenced by unusual occurrences and to disregard the ordinary happenings. In any sort of investigation, statistical or otherwise, the unusual cases must not exert undue influence. Many people are of the opinion that to break a mirror brings bad luck. Having broken a mirror, a person is apt to be on the lookout for the expected "bad luck" and to attribute any untoward

event to the breaking of the mirror. If nothing happens after the mirror has been broken, there is nothing to remember and this result (perhaps the usual result) is disregarded. If bad luck occurs, it is so unusual that it is remembered, and consequently the belief is reinforced. The scientific procedure would include all happenings following the breaking of the mirror, and would compare the "resulting" bad luck to the amount of bad luck occurring when a mirror has not been broken.

Statistics, then, must include in its analysis all sorts of happenings. If we are studying the duration of cases of scarlet fever, we may study what is typical by determining the average length and possibly also the divergence below and above this average. When considering a time series showing steel-mill activity, we may give attention to the typical seasonal pattern of the series, to the growth factor (trend) present, and to the cyclical behaviour. Sometimes it is found that two sets of statistical data tend to be associated and it behooves us to ascertain what is typical in the relationship. In the chapter on correlation it is pointed out (p. 651) that there is an association between temperature and the rapidity with which crickets chirp. If the temperature increases, the crickets chirp faster; if the temperature decreases, the crickets chirp more slowly. The relationship can be expressed mathematically and we can estimate the rapidity of crickets' chirps from the temperature; or, conversely, lacking a thermometer, we can make a good estimate of the temperature based upon the rapidity of chirps.

Occasionally a statistical investigation may be exhaustive and include all possible occurrences. More frequently, however, it is necessary to study a smaller group or sample. If we desire to study the expenditures of lawyers for life insurance, it would hardly be possible to include all lawyers in the United States. Resort must be had to a sample; and it is essential that the sample be as nearly representative as possible of the entire group, so that we may be able to make a reasonable inference as to the results to be expected for an entire population. The problem of selecting a sample is discussed in the following chapter. In Chapters XII and XIII an attempt is made to determine how much reliance may be placed in the results obtained from a sample. We should also have some idea what variation in results to expect if additional samples are selected. These are important considerations, since an unrepresentative sample or an unreliable statistical measure may cause us to draw false and unwarranted conclusions.

Sometimes the statistician is faced with the task of forecasting. He may be required to prognosticate the sales of automobile tires a year hence, or to forecast the population some years in advance. Several years ago a student appeared in a summer session class of one of the writers and in

a private talk announced that he had come to the course for a single purpose: to get a formula which would enable him to forecast the price of cotton. It was important to him and to his employers to have some advance information on cotton prices, since the concern purchased enormous quantities of cotton. Regrettably, the young man had to be disillusioned. To our knowledge, there are no magic formulae for forecasting. This does not mean that forecasting is impossible; rather it means that forecasting is a complicated process of which a formula is but a small part. And forecasting is uncertain and dangerous. To attempt to say what will happen in the future requires a thorough grasp of the subject to be forecast, up-to-the-minute knowledge of developments in allied fields, and recognition of the limitations of any mechanical forecasting device. Further comments concerning forecasting are to be found in Chapter XXV.

Interpretation. The final step in an investigation consists of interpreting the data which have been obtained. What are the conclusions growing out of the analysis? What do the figures tell us that is new, that reinforces or casts doubt upon previous hypotheses, or (if the study is sufficiently inclusive) that proves or disproves former beliefs? The results must be interpreted in the light of the limitations of the original material. Too exact conclusions must not be drawn from data which themselves are but approximations. It is essential, however, that the investigator discover and clarify all the useful or applicable meaning which is present in his data.

A Few Improprieties

The research worker must be constantly on the alert to avoid any misuses of his material. Illogical and careless reasoning or improper use of data will destroy the value of a study which may be technically acceptable in its earlier phases. A few examples of fallacious procedures may clarify this point. In later chapters of the book, other fallacies are occasionally mentioned in connection with the methods to which they apply.

Bias. The presence of bias on the part of an investigator is, obviously, sufficient to discredit the entire undertaking. Bias may be conscious or deliberate; in such a case it is synonymous with falsification. On the other hand, an unconscious bias may be operative, and this, perhaps, is a more dangerous form since the analyst himself may not be aware of it. The following is an illustration of apparently unconscious bias:[1]

> A friend had invited an acquaintance to lunch, and found at the end
> of the meal that he had left his purse in the office and had no money.
> The acquaintance, at his request for a loan, took out a five-dollar bill

---

[1] From "The Mind of a Child" by Jessica Cosgrove, *Good Housekeeping*, January 1927, p. 206.

and a ten-dollar bill. My friend took one of them—to this day he does not know which—telling his acquaintance not to let him forget the loan. He did forget it, however, until several weeks later when they met again, and each wrote on a piece of paper the sum he thought had been borrowed. The lender wrote ten, and the borrower five. They were both psychologists, so each searched his memory carefully, and each had circumstantial evidence that seemed to each conclusive, to prove himself right. Neither cared about or needed the money especially, but to them it indicated a universal principle, that each of us interprets and remembers facts in the form most agreeable to himself. No wonder both sides must be represented in courts of law, and that much honestly given evidence must be rejected!

As will be seen in the following chapter, statistical data cannot be picked out of thin air as the conjurer appears to produce coins at his finger tips. The process is one requiring care and attention to details. The data, when obtained, should be of value and not be casually disregarded. Note what a reviewer said of a certain author:

> Blank is thorough and undaunted. Have statistics on any subject been collected before? He has collected more and better ones. If it is by its intrinsic nature unchartable, he has charted it none the less. . . . Chronology itself fares badly in his hands at times. If his examples require to be a century or two misplaced, Blank can forget even his statistics and his charts in the good cause of logic.

Omission of important factor. Shortly after the introduction of the all-metal top for automobiles, a certain manufacturing company felt called upon to prove that all-metal tops did not result in hotter car interiors. They suggested a test involving three steps:

1. Take a piece of top fabric about 8 inches square. Place a piece of lining material of similar size beneath the fabric, and a thermometer beneath the lining material.

2. Take a piece of highly finished steel about 8 inches square. Place similar sized pieces of $\frac{1}{4}$-inch felt and lining material beneath the metal, and a thermometer beneath the lining material.

3. Place each of the above assemblies on a board at room temperature. Carry the entire apparatus out into hot sunshine, leave it exposed for about 10 minutes, and then read the temperatures of the two thermometers.

The difficulty with the above experiment is that the reader is asked, in step 2, to use a piece of *highly finished* steel. Automobile tops are painted—many of them with black or a dark color of paint—and there-

fore absorb more heat than does *highly finished* steel. The obvious fallacy in the test vitiates the experiment, although the additional insulation may actually make the metal-top car cooler than the fabric-top car.

Carelessness. We cannot go through life without making mistakes, but carelessness should be reduced to a minimum. The wife of one of the authors wrote to a large department store to ask the size of a cedarized storage chest. The reply said, "This merchandise is available in the $3'' \times 1'' \times 1\frac{1}{2}''$ size."

Many of us have received sealed envelopes minus enclosures, or postal cards blank on the message side, and have, perchance, been guilty of sending the grocer's bill back to the grocer minus the check or with the check unsigned.

A study of salaries was under way and a certain corporation had been requested to furnish data concerning its employees. A note to its report appeared substantially as follows: "All salaries under $5,000 per annum are shown as the maximum for each type of work. The assistant to the auditor stated that the maximum is equivalent to a general average for each group." Perhaps this is an illustration of a conscious bias on the part of the assistant to the auditor. It must be obvious that, if the maximum and the average are the same, then there are no values below the maximum.

Non-sequitur. A weekly news magazine, the circulation of which had been growing in a healthy fashion, undertook in 1936 to demonstrate that its readers greatly exceeded its circulation. After showing figures of its circulation, the magazine stated: "And each of these subscribers represents 3.26 cover-to-cover readers, according to former Deputy Police Commissioner ———, who counted and identified [sic] 216,948 fingerprints on copies his operatives had picked up at random from subscribers' homes in seven different cities or towns." How could the investigator *know* the fingerprints belonged to cover-to-cover readers? Or, did he find each fingerprint on *every* page and, if so, does that prove each page was read? Do you ever actually read a magazine from cover-to-cover?

Non-comparable data. In July 1936, newspapers carried reports of a meeting of the American College of Osteopathic Obstetricians at which a doctor is reported, by a metropolitan paper, to have stated that the maternal death rate among mothers treated by osteopathic physicians is less than half that among cases handled by the medical profession. The higher rate in the latter instance was said to be due to excessive use of anaesthetics, interruption of labor, and undue reliance on mechanical devices. A survey of 14,000 osteopathic delivery cases was said to show a maternal death rate of 2 8 per thousand cases. This figure was compared with the nation's average of more than 6 per thousand. It should

be obvious that the average rate for the entire country is not representative of the rate for cases attended by the medical profession, since many maternity cases are not attended by physicians.

The makers of a small, inexpensive car had been stressing the fact that the introduction of their car had converted many used-car buyers into new-car owners. Concerning costs of operation, they pointed out that "owners report up to thirty-five miles to the gallon of gasoline, which compared with the average mileage obtained with a used car . . . is a saving of great importance to persons in the low-income group." The comparison of *maximum* mileage for one type of car with *average* mileage for other types of used cars is certainly unjustified.

Confusion of association and causation. Sometimes factors which are associated are erroneously regarded as being causally related. A southern meteorologist discovered that the fall price of corn is inversely related to the severity of hay fever cases. This does not imply that the low price of corn causes hay fever to be severe, nor does it imply that severe cases of hay fever bring about a drop in the price of corn. The price of corn is generally low when the corn crop has been large. When the weather conditions have been favorable for a bumper corn crop, they have also been favorable for a bumper crop of ragweed. Thus the fall price of corn and the suffering of hay fever patients may each be traced (at least partly) to the weather, but are not directly dependent upon each other. A further discussion of association and causation is given in Chapter XXII.

Another instance of the confusion of association with causation is illustrated by Chart 1. In connection with this chart it was asserted, "When farm income goes up, factory payrolls invariably follow, but they do not lead the procession. One is cause, the other effect." If such a procession does exist, it can hardly be shown by annual data. If factory payrolls *follow* farm income, we should show that fact by plotting monthly data as is done for two other series in Chart 253, page 807. As to the causal relationship, it is fairly obvious that, while an increase (or decrease) in farm income does have a corresponding effect upon factory payrolls, the payrolls in turn have a reciprocal effect upon farm income. Furthermore, both are dependent upon any other factors which tend to affect the pattern of general business.

Insufficient data. Insufficient data result in a high degree of uncertainty respecting any conclusion which may be made from them. A very small sample may lead us to a correct conclusion, but we cannot be sure of our conclusion. When a physician is developing a new treatment, he does not announce its efficacy after trying it out on a few individuals. He

Dependence of Factory Pay Rolls on Farm Income

All primary purchasing power originates in the soil, except that obtained from fisheries and mines. When farm income goes up, factory payrolls invariably follow, but they do not lead the procession. One is cause, the other effect.

It is high time for American Industry to forge an alliance with American Agriculture using the tools of modern Science which enable a wide variety of farm products to be diverted to industrial rather than to food uses.

In proportion as this is done the problem of surplus crops and of ruinous prices will disappear, farm purchasing power will be augmented and stabilized, the demand for industrial products will steadily increase, the total employment of all men and women able and willing to work will be required, the drain on the public treasury for unemployment relief will subside economic security will replace industrial strife and self-determination both of Agriculture and of Industry will be restored.

CAUSE AND EFFECT

GROSS FARM INCOME

TOTAL FACTORY PAYROLLS

SOURCE OF DATA
U S BUREAU OF AGRICULTURAL ECONOMICS,
U S. BUREAU OF LABOR STATISTICS.

Goal

YEAR: 1921 1922 1923 1924 1925 1926 1927 1928 1929 1930 1931 1932 1933 1934 1935 1936 1937 YEAR

$ 1,000,000,000 2,000,000,000 3,000,000,000 4,000,000,000 5,000,000,000 6,000,000,000 7,000,000,000 8,000,000,000 9,000,000,000 10,000,000,000 11,000,000,000 12,000,000,000 13,000,000,000 14,000,000,000 15,000,000,000

Chart 1. Gross Farm Income and Total Factory Payrolls, 1921-1936. In the original chart gross farm income was shown by red bars and total factory payrolls by black bars. (From the National Farm Chemurgic Council)

11

must have enough data to be relatively sure of results. If two or three subjects respond favorably, he cannot be sure that the occurrences were not due to chance. The favorable responses of these few might have come without the treatment, or in spite of it! Of course, there must be a "control" group to show how the subjects would respond without any treatment. Moreover, both the control group and the treated group must be sufficiently large to warrant a conclusion. A discussion concerning the size of a sample and the reliability of values computed from it is given in Chapters XII and XIII.

Unrepresentative data. Conclusions may be based upon data which are numerically sufficient, but which are not representative. A small sample *may* be representative; on the other hand, a large sample *may not* be representative.

An example of a conclusion based upon unrepresentative data is the forecast of the 1936 presidential election as made by the *Literary Digest*. More than 10,000,000 straw ballots were sent out by the *Digest*. Of these, 2,376,523 were returned and they indicated that 370 electoral votes would be cast for Landon and 161 for Roosevelt. The final election results were 523 electoral votes for Roosevelt and 8 for Landon. The difficulty was that the mailing lists used as a basis for the poll were relatively heavily weighted with persons in the upper economic brackets and thus were not representative of the entire voting population.

Concealed classification. Conclusions drawn from statistical data may sometimes be invalid because of the presence of a concealed classification which is overlooked. The fallacy of concealed classification is illustrated by some data appearing in the *Monthly Labor Review* for February 1938 and concerning which its readers were warned. Data were presented showing the union wage rates in Hebrew and in non-Hebrew bakeries. It appeared from the figures that Hebrew bakeries paid an average hourly rate about 50 per cent higher than non-Hebrew bakeries. Qualifying this, the *Review* said, "Although Hebrew bakeries generally have higher rates, one reason for this large difference is the fact that a large proportion of the Hebrew bakeries are located in New York City, where the average of all rates is higher than in other localities."

A concealed classification was found to be present in a study of suicides. The data seemed to show that suicides were more likely to occur among certain religious groups than among others. Upon further consideration it was apparent that the matter of the urban or rural occurrence of the suicides had been overlooked. Hence the conclusion should have been— not that suicides tended to tie up with given religious groups—but that suicides were more common in urban territories and that these religious groups were also more numerous in the cities.

Research Methods

It must not be assumed that the statistical method is the only method to use in research; neither should this method be considered the best attack for every problem. Just as the carpenter has a number of tools, each appropriate for a different sort of operation, so the researcher can avail himself of various techniques which are the tools of his trade and each of which is appropriate to a specific type of situation. If an amateur carpenter uses a screwdriver in lieu of a chisel, the results are not likely to be either workmanlike or satisfactory. Similarly, it is important that the investigator consider his problem carefully at the outset and make use of the technique or techniques which are appropriate to it. Just as the carpenter needs to use more than one tool in completing a piece of work, so the research worker must often make use of, not one, but several methods.[2]

When we desire a great deal of information concerning each individual or occurrence to be studied, much of our data may be non-quantitative by its very nature. In such an event we employ the *case method* of investigation, the purpose of which is to study in detail the characteristics peculiar to the individual case and to generalize from a number of such detailed studies. Some of the information obtained in a case study (such as wages, number of offspring, etc.) may be statistical and when many cases are included, statistical summaries may be made of the non-quantitative information obtained.

Sometimes a problem may be attacked by the historical approach. Although the *historical method* is largely descriptive and non-quantitative, we may find statistical aspects when we consider growth or decline of imports, exports, population, and other series.

Again, the appropriate procedure may be to make use of the *experimental method*, in which we allow only the factor we are studying to vary, and thus we attempt to control as many as possible of the other factors. For example, if we wished to study the effect of car weight upon tire mileage, we should control road conditions, speed, temperature, size of tire, quality of rubber and of cord, inflation of tire, and many other factors.

In the social sciences, the experimental method can rarely be applied and certain aspects of the statistical method are used in lieu of it. We cannot, for example, ascertain the effect of different sorts of diets upon length of life, by forcing groups of people to live upon prescribed diets and by actually making all other phases of their lives identical. Instead,

---

[2] Various methods are described in Manuel Conrad Elmer, *Social Research*, Prentice-Hall, Inc., New York, 1939, and in Walter Earl Spahr and Rinehart John Swenson, *Methods and Status of Scientific Research*, Harper and Brothers, New York, 1930.

we must find groups of people on different diets, and then we must measure the importance of and control statistically as many as possible of the other phases of their lives since we cannot control them experimentally. The experimental and statistical methods are not antithetical, but under practical conditions the statistical method supplements the experimental method. If an experiment could be so designed that *all* variables were *completely* controlled, statistics might not be needed. At best we can usually control but a few of the more important factors, and thus it is necessary to evaluate statistically the importance of a host of other minor disturbing factors (sometimes designated as "chance"), as described in Chapters XII and XIII.

Some problems may be approached by the *deductive method* rather than by the *inductive method*. When a hypothesis has been set up deductively and when quantitative data are available, statistics may enable an inductive test to be made of the hypothesis, and this test may serve to support or discredit the hypothesis. Conversely, relationships arrived at statistically (as, for example, the rather close negative association found in some states concerning the size of farms and the value of land per acre) may suggest causal connections which may be worked out deductively. Again we have two methods which are not antagonistic, but complementary.

Selected References

R. E. Chaddock· *Principles and Methods of Statistics*, Chapters I, II, III; Houghton Mifflin Co., Boston, 1925. Chapter II contains illustrations of misuses of statistical data.

F. E. Croxton and D. J. Cowden: *Practical Business Statistics*, Chapter I; Prentice-Hall, Inc , New York, 1934. From the point of view of business statistics. Gives additional illustrations of misuses of statistical data.

M. C. Elmer: *Social Research*, Chapters I-XV; Prentice-Hall, Inc., New York, 1939. Methods of research are discussed.

G. A. Lundberg: *Social Research*, Chapters IV, VIII; Longmans, Green and Co., New York, 1929. Methods of research.

W. E. Spahr and R. J. Swenson: *Methods and Status of Scientific Research*, Chapter X; Harper and Brothers, New York, 1930. Research methods

P. V. Young: *Scientific Social Surveys and Research*, Chapters V, IX, X, XIV; Prentice-Hall, Inc., New York, 1939. Methods of research.

CHAPTER II

STATISTICAL DATA

When a research worker desires statistical data concerning a topic of interest, it may be that he can choose between collecting the data himself or obtaining the needed figures from published or unpublished compilations. If an individual or organization has prepared reliable data which are pertinent to the problem, it is vastly less expensive to make use of the existing information. Although to collect one's own data is more costly, the procedure enables the investigator to obtain exactly that information which is needed to answer the specific questions that are under consideration.

Not all readers will be faced with the problem of collecting original statistical data; many will find it possible to refer to existing sources for information. The data from such sources may be evaluated and more intelligent use may be made of them if the research worker has some knowledge of the procedure and pitfalls involved in collecting, editing, and marshalling statistical data.

An illustration cited by Stamp[1] is to the point: Harold Cox, when a young man in India, quoted some Indian statistics to a judge. The judge replied, "Cox, when you are a bit older, you will not quote Indian statistics with that assurance. The government are very keen on amassing statistics—they collect them, add them, raise them to the nth power, take the cube root and prepare wonderful diagrams. But what you must never forget is that every one of those figures comes in the first instance from the *chowty dar* (village watchman), who just puts down what he damn pleases." It should be added that this story refers to the India of a day long past Today India has many able statisticians and an active statistical society. Presumably the *chowty dar* no longer functions as the source of local statistical information.

The process of collecting statistical data will be examined first in the

---

[1] Sir Josiah Stamp, *Some Economic Factors in Modern Life*, pp. 258-259, P. S. King and Son. London, 1929.

15

following text.   Later in the chapter, attention will be directed toward the use of statistical sources.

Collecting Statistical Data

Method of collection.   Statistical data are frequently obtained by a process in which the desired information is obtained from the house-holder, business man, or other informant, either by an enumerator who visits the informant and asks the necessary questions (entering the replies on a schedule), or by sending to the informant a list of questions (some-times called a *questionnaire*) which he may answer at his convenience.   The data collected at each population census are obtained by the *enumeration process*.   Sometimes information is obtained by *registration*, which means that the information is reported to the proper authority when (or shortly after) an event occurs.   Thus births and deaths must be registered.   In many states automobile accidents must be reported to the commissioner of motor vehicles.

In general outline the problems of obtaining data by mailing question-naires, by enumeration, and by registration are similar.   Under a system of registration there is, of course, the difficulty that many persons will neglect to register.   Constant vigilance and frequent checkups are neces-sary on the part of the registrar.   Registration, however, is usually with a properly designated government official, and there is ordinarily legal compulsion that the data be supplied.   Since most statistical information is obtained by mailing questionnaires or by enumeration, the balance of this section will be devoted to the procedure for collecting data by such methods.

Outline of procedure.   The steps in a statistical investigation may be designated as follows:

1.   Laying out the general plan.
2.   Devising questions and making the schedule.
3.   Selecting the sample (if the enumeration is not to be a complete one).
4.   Using the schedules to collect the data.
5.   Editing the schedules.
6.   Tabulating the data.
7.   Preparing finished tables and charts.
8.   Analysis and interpretation.

The procedure will usually be in the order listed above, except that the selection of the sample may be handled as a phase of the first step, laying out the general plan.

Laying out the general plan.   If a topic is to be studied statistically, it behooves the investigator to become familiar at the outset with what has

already been done by others.   He may find that someone else has already examined the same topic and that his questions have already been answered.   He may wish to design his study so that it can be compared with those which have preceded his.   He will doubtless profit by the experience and the mistakes of others.   He may find that the difficulties involved in the investigation of his topic are so great that they are insurmountable; the cost may be too great, or it may appear that informants do not wish to divulge the type of information which is needed.

Having studied what has been done by others, the investigator is ready to consider the general aspects of what he would like to know.   If an unemployment study is projected, there are many inquiries concerning each individual which are pertinent.   The following suggests some of the more important ones:

> Does the individual have any dependents?   How many?
>
> Is the person male or female?
>
> Is he married?
>
> How old is the person?
>
> Is he native white, native colored, or foreign born?   If foreign born, from what country?
>
> Does he own property?
>
> What is his usual occupation?   In what industry?
>
> What type of work is he doing at present?   (If the study is a detailed one, consideration may be given to listing the job experience of the individual for a number of years, together with the wages received.)
>
> Is he employed full time?   Part time; if so, what fraction?   Is he entirely unemployed?
>
> If the individual is working part time or is totally unemployed, what is the reason?
>
> If he is totally unemployed, how long has he been so?   Also, is he able to work and willing to work; or, alternately, is he actively looking for work?

The reader will doubtless think of other questions of importance, but these suffice to indicate the nature of this preliminary step.   Usually we cannot undertake to obtain answers to all the questions which are important.   It may be too expensive to make so comprehensive an inquiry. There may be some questions (such as the one in regard to property ownership, and the one in regard to wages) which informants will often decline to answer.   The most important and practicable questions are therefore selected to form the basis of the inquiry.   It is these which will be incorporated into the schedule.

There are several matters of general importance which are often considered in connection with laying out the general plan.   One of these has

to do with the extensiveness of the study. Will it include the entire community or merely a sample? If funds and enumerators are available, we may make a complete enumeration; often we must be satisfied with a sample. We shall consider the selection of the sample after we have completed the discussion of the schedule.

Another problem concerns whether the schedule is to be sent out by mail (in which case it must be very simple and self-explanatory) or whether enumerators are to be used. If use is to be made of paid enumerators, it is necessary to locate qualified persons. However, it is often true that funds are not available to hire enumerators. In fact, it is sometimes the case that, valuable as the results of an investigation might be, they are not worth what it would cost to employ enumerators! Studies have been made using, as unpaid enumerators, policemen, college students, postmen, truant officers, and even school children.

A third matter has to do with the place where the informants will be interviewed. In the case of the unemployment study we could send enumerators to interview people at their work, in the streets, or at home. It is obvious that the last of the three is preferable. For the unemployment study we should also consider whether to list on our schedule all the people in a household, irrespective of age, sex, desire for work, and mental or physical condition. To list everyone would give us a complete picture, but it would also clutter up the schedule with relatively useless information. For the purposes of an unemployment study we are ordinarily not interested in housewives who seek no work outside the home or in young children. We may be interested in elderly men, in an attempt to learn what proportion of the population is retired or is considered too old or infirm to work. Thus it may be desirable to exclude all persons below (say) 16 or 18 years of age, and all females not usually employed.

Devising questions and making the schedule. It has already been pointed out that not all the questions which we would like to have answered can be included in the schedule. Having selected those points which we wish to include in our inquiry, we must formulate each question so that it may be readily and accurately answered, and then we must draft the schedule form. The accompanying schedule form used in unemployment studies in Buffalo, New York, shows how some of the questions concerning unemployment may be worked into a very simple schedule form.[2] Shown

---

[2] From Frederick E. Croxton, *Unemployment in Buffalo, November 1932*, Special Bulletin No. 179, Division of Statistics and Information, New York State Department of Labor.

The Buffalo study will be referred to frequently to illustrate various points in this chapter. No inference should be drawn from this that it is considered a model study. It is, however, simple enough in its general outlines to facilitate explanation of numerous principles and methods.

Address _____ Territory _____ Visitor _____

| Relation to head of household 1. | Sex 2. | Age 3. | Nativity 4. | Present or last regular employment | | | Employed now | | Unemployed now | | | |
				Employer 5	Industry 6.	Occupation 7	Full time (√) 8.	Part time fraction 9	Weeks unemployed 10	Reason for unemployment 11.	Able to work 12	Wants to work 13
a. Head												
b.												
c.												
d.												
e.												
f.												
g.												

Notes _____

N. Y. State Dept. of Labor and The Buffalo Foundation   Unemployment Survey, Buffalo, N. Y. 1932

19

Schedule used in the Buffalo unemployment studies, 1930-1932. The forms were printed on cards 5 × 8 inches in size

Form 15-4

DEPARTMENT OF COMMERCE—BUREAU OF THE CENSUS

FIFTEENTH CENSUS OF THE UNITED STATES, 1930

POPULATION SCHEDULE

Sheet No _____

A

Enumeration District No _____

Supervisor s District No _____

State _____

County _____

Township or other division of county _____

Incorporated place _____

Ward of city _____

Block No _____

Unincorporated place _____

Enumerated by me on _____ 1930 _____ Enumerator

| PLACE OF ABODE | NAME | RELATION | HOME DATA | PERSONAL DESCRIPTION | EDUCATION | | | MOTHER TONGUE, OR NATIVE LANGUAGE OF FOREIGN BORN | CITIZENSHIP, ETC. | OCCUPATION AND INDUSTRY | | | EMPLOYMENT | VETERANS | |

ABBREVIATIONS TO BE USED AT COLUMNS INDICATED

ENTRIES ARE REQUIRED IN THE SEVERAL COLUMNS AS FOLLOWS

20

also is the schedule used in the 1930 Census of Population. It so happens that both of these inquiry forms are set up in columnar arrangement with a line for each individual. Schedules are not necessarily made in this fashion. The schedule used for a study of accidents is merely a carefully arranged series of questions; so also is the form used in one of the inquiries made by Hartwell, Jobson, and Kibbee concerning food. The question form used in a mail inquiry must be even easier to understand than the schedule used by enumerators. Business houses find it pays to send out questionnaires that are attractive and interesting, and that require a minimum of effort in answering, for by so doing they receive a larger proportion of replies from their mailing lists.

Schedule Used by the United States Bureau of Labor Statistics in a Study of Industrial Accidents.

Observe that there are notations to assist the enumerators at the bottom of the population schedule. Instructions were given in a separate booklet, in which 33 pages were devoted to the population schedule. A separate sheet of instructions was furnished to the enumerators in the Buffalo unemployment study.

The construction of statistical schedules is something which is learned most satisfactorily by actually making and using them. Nevertheless, there are some cautions which are helpful:

1. *Clarity is essential.* The entire schedule as well as each question should be as simple and as clear as possible. This is particularly true of schedules (sometimes called questionnaires) sent to, or left with, persons

to be filled out at their convenience. An ambiguous question or a question that invites an ambiguous answer produces useless data and involves wasted time and money. An organization, in making a study, queried some hundreds of parents: "Is your child's outlook on life broader or narrower than yours was at the same age?" The investigator presumably expected the replies to read "Broader" or "Narrower." Replies actually

---

Int___ ___Date___Neighborhood___

Com___Econ___Hse___Tel___Ref___Car___

Serv___Sex___Col___Age___Occ___Agr___

FOOD INDUSTRIES

1. Have you tried frozen foods?___

2a. If yes—   2b. If no—

   Were they satisfactory? Yes ( ) No ( )   Why not?

   ___

   ___ ___

3. What canned foods are most nearly as good as when fresh?___

   ___

4. Is it dangerous to leave food overnight in an opened can? Yes ( ) No ( )

5. Do you prefer milk in bottles ( ) or cartons ( )?

6. How do you prefer to buy fresh coffee—

   In sealed tins . .. . ( )
   In dated packages . ... ( )
   Ground when bought ..... ( )
   In the bean for home grinding ( )

Schedule Used by Hartwell, Jobson, and Kibbee (Public Relations Counsel) to Obtain Data for a Food Industry Trade Journal. Data are collected entirely by interviewers. Entries following the abbreviations at the top of the schedule are to assist in selecting a sample which will properly represent all relevant strata of the population. The meanings of the abbreviations are as follows· Int, name of interviewer; Com, size of community; Econ, economic group; Hse, type of house; Tel, telephone; Ref, refrigerator; Serv, servants; Col, color or race; Occ, occupation: Agr, agreeableness of person interviewed.

received, however, were frequently "Yes," "No," "I doubt it," and "I hope so"—none of which had any meaning. Furthermore the question is so worded as not to allow for the fact that there may be two or more children in the family. The inquiry concerning marital condition when put "Married or Single?" is open to two objections: (1) Either a "Yes" or "No" answer is meaningless; (2) not all persons are included in these two categories. One good way of asking this question is to say:

Check whether:

   Single...............
   Married..........
   Widowed.........
   Divorced.........
   Separated......

The investigator should not be satisfied merely with wording his questions so that they can be understood; he should draft them so carefully that they cannot be misunderstood.

2. *Not all questions can be accurately answered.* No matter how clearly a question is stated, there are some sorts of inquiries which are apt to elicit unsatisfactory returns. The schedule of the United States Census of Population asks for age at last birthday for each person enumerated. Reference to the 1930 Population Volume II, Tables 20 and 21 and chart on page 571, shows a peculiar distribution of the population by one year age groups. Beginning with age 30 and continuing through age 80, there are definite concentrations of persons on every age ending in 0 or 5. For example, there are *more* people reported as 35 than there are as either 34 or 36. There are secondary concentrations upon certain ages which are a multiple of 2, most noticeable when these even numbers of years are not adjacent to an age ending in 5. Thus there are concentrations at 28, 32, 38, 42, 48 and so on through 72. Furthermore, there seem to be too many males aged 21 and too many females aged 18. In its instructions to enumerators the Census warns that many persons will report age in round numbers and says, "Therefore, when an age ending in '0' or '5' is reported, you should inquire whether it is the exact age. If, however, it is impossible to get the exact age, enter the approximate age rather than return the age as unknown."

The rounding of ages is not peculiar to the Census. Some of the factors believed to lead to reporting ages in round numbers are: (1) The information concerning an individual is not necessarily furnished to the enumerator by the person himself; it is often given by a relative, friend, landlady, or other person, and some of these informants cannot have exact information.

(2) When ages are intentionally misstated, as they occasionally are, there is reason for believing that they are often rounded. (3) Some persons are careless, or occasionally a person of low intelligence may always think in terms of round numbers. The Census notes that the rounding is most noticeable for those classes of the population in which the proportion of illiterates is greatest. (4) A few persons do not know their exact ages. (5) There may be carelessness on the part of enumerators. Some improvement in the accuracy of reporting ages may be had by asking date of birth instead of, or in addition to, age. It should be recognized, however, that the posing of a more exact question does not produce better data when exact knowledge is lacking, as in the case of a landlady reporting for her roomers. Furthermore, the matter of the expense involved in asking this additional question might more than offset the expected increase in accuracy. When age is of primary importance, as in the case of application for insurance, date of birth is usually asked and may be verified by documentary evidence.

Another interesting example of thinking in terms of round numbers occurred in the case of a contest sponsored by a motion picture theatre. An irregular-shaped glass jar was filled with cranberries and six prizes were offered to the patrons who guessed most nearly the correct number of cranberries in the jar. An analysis of the 1,996 guesses showed that there were 1,465 which ended in 0 or 5.

3. *Certain types of questions should be avoided.* When the prosecuting attorney asked the alleged wife beater, "Have you stopped beating your wife?" he attempted to put the defendant, whether he replied "Yes" or "No," in the position of admitting that he had beaten his wife. In a scientific investigation we should scrupulously avoid leading questions. When asking the reason for unemployment in 1932, an enumerator would have been suggesting the answer if he had said, "I suppose you are unemployed because of the depression?" Rather he should have inquired, "What is the reason you are unemployed?"

Questions which are unduly inquisitive or which are liable to offend should likewise be avoided. In a study of social workers each married woman was asked whether or not she lived with her husband. The inquiry was injudicious, aroused resentment, and would hardly have been productive of useful data if it had been answered by all the persons queried. Questions concerning personal matters (such as income) should be handled with tact—perhaps asked at the close of the interview after the cooperation of the informant has been secured. Sometimes it is better not to ask such a question but to infer the general income level from knowing if there is a telephone in the home; if the home is owned, and its apparent value; the

wage earner's occupation; make of car(s) driven, if any; servants employed, if any; etc.

4. *Answers should be objective and capable of tabulation.* When making factual studies, questions should be so designed that objective answers will be forthcoming. Instead of asking the condition of a building and allowing the enumerator to state the condition in his own words, the Real Property Inventory (United States Department of Commerce, 1934) asked if a structure was in good condition, needed minor repairs, needed structural repairs, or was unfit for use. Although the answers to these questions are not completely objective, at least they are capable of being readily tabulated.

5. *Instructions and definitions should be concise.* The enumerator and informant should never be in doubt as to what information is desired and what terms or units are to be used. When inquiring as to the employment status of an individual (whether full time, part time, or unemployed), our inquiry must be as of some specific time. Thus the 1930 Unemployment Census of the United States considered as unemployed those gainful workers who were not at work on the day preceding the visit of the enumerator (or on the last previous work day, in case that day was not a regular working day for the person enumerated). The 1932 Buffalo unemployment study asked employment status as of November 4, 1932.

If information is desired as to the exact situation of a part time worker, it must be made clear whether the desired answer should be: (1) hours per day; (2) hours (or days) per week; or (3) fraction of usual full time.

The units used in a study should be clearly understood by both the enumerator and the informant. If we are collecting data on coal production or consumption, we should state clearly whether we are referring to short or long tons. If we desire information as to the number of rooms in houses, it should be clearly understood whether or not bathrooms, kitchenettes, powder rooms, dressing rooms, and the like are to be counted as rooms.

6. *Arrangement of questions should be carefully planned.* Not only must the questions be well arranged on the schedule form to allow proper space for answers, but the arrangement of the questions should be such as to facilitate the answering of each question in turn. If a logical flow of thought is involved, it should be followed in the arrangement of questions. Questions should not skip back and forth from one topic to another.

After a schedule has been drafted, the desirable procedure is to try it out with a group, discover its shortcomings, and then revise it in the light of the tryout. If there is not time for a tryout, ask some competent investigators to go over it and make suggestions for its improvement. When

the final form of the schedule has been decided upon, careful instructions for filling it out should be prepared. If the schedules are to be mailed to the persons furnishing information, these directions should be as clear and concise as possible. If enumerators are used, the instructions to the enumerators should be complete in order to cover as many as possible of the situations which may occur in their work

Selecting the sample. As pointed out previously, the United States Population Census is a complete enumeration of the people of the United States. By this we do not mean that not even one individual is omitted, because it is, of course, true that a few persons are not enumerated. However, the intent is to include everyone, and the very few who are not included are likely to be those in extremely out-of-the-way places, traveling men with no permanent abode, tramps, etc. Similarly the Census of Agriculture undertakes to include all farms in the United States, and the Census of Unemployment (1930) attempted to embrace all unemployed persons.

Sometimes it is not practicable or necessary to essay a complete enumeration. We may be satisfied to have an almost complete coverage. Thus the United States Census of Manufactures does not undertake to include all manufacturing establishments but eliminates the extremely small concerns. The quinquennial manufacturing censuses from 1904 to 1919 included factories having products valued at $500 or more during the calendar year. Beginning in 1921, censuses of manufacturing were taken every two years and, at these biennial censuses, data were collected from only those establishments having products valued at $5,000 or more. The importance of this exclusion was studied in 1921 (when certain general data were obtained from the smaller establishments) and it appeared that, while the firms having products valued at $500 to $5,000 constituted 21 per cent of the total number of establishments, they employed only six-tenths of one per cent of the total number of wage earners and had an output of but three-tenths of one per cent of the total value of products. Thus the enumeration was quite incomplete in respect to number of establishments, but virtually all-inclusive in regard to number of wage earners and value of products.

Although the Census of Agriculture undertakes to include all farms, it does not, therefore, include all land used for agricultural purposes. A farm, as defined by the census, is: "All the land which is directly farmed by one person conducting agricultural operations either by his own labor or with the assistance of members of his household or hired employees" But, enumerators are warned, "Do not report as a 'farm' any tract of land of less than 3 acres, unless agricultural products to the value of $250 or

more were produced on such tract in 1929." [3] It is thus apparent that very small plots used for agricultural purposes are not included, but the coverage of agricultural lands is, nevertheless, virtually complete.

It may be too expensive or too time-consuming to attempt either a complete or nearly complete coverage in a statistical study. Furthermore, to arrive at valid conclusions, it may not be *necessary* to enumerate all or nearly all of a population. We may study a sample drawn from the larger population and, if that sample is adequately representative of the population, we should be able to arrive at valid conclusions. There are various ways in which a sample may be selected from a population. No matter which of these is employed, it must be remembered that the cardinal purpose is to obtain a representative sample, that is, one which contains all elements in the same proportion as in the population from which it is drawn. In short, it is *not* merely a matter of grabbing *any* 2, 5, 10, or 20 per cent sample of a population, but of selecting that sample in such a way that it will be as representative as possible.

1. *Random sample.* One method of selecting the items to comprise a sample consists of drawing them at random. More exactly, the items should be drawn independently so that each item will have an equal chance of being selected. When this situation holds, it is more likely that the sample will have the different elements in the same proportion that they exist in the population, than that these elements will be present in any other proportion. Such a situation may be rather exactly realized in drawing marbles from a large container (which holds, say, several thousand marbles, $\frac{1}{3}$ of which are white, $\frac{1}{3}$ black, and $\frac{1}{3}$ red) if we draw one marble at a time, replacing it after each draw and thoroughly mixing the marbles before each draw. It may be closely realized as we draw samples of screws, nails, bricks, wire, or other products from the production stream of a factory. It may be approximately realized in a community study, but only approximately so because of the difficulty of setting up a selection procedure. If the selection of a sample is to be based upon individuals or households, it is necessary to have a listing of those persons or households so that the sample may be selected. Sometimes a city directory for individuals, or the list of subscribers for electricity, gas, or water for households, may serve as a basis, and every tenth or twentieth name (depending upon the size of the sample desired) may be selected. Lists such as these are obviously incomplete, and sometimes selectively so, in that certain categories of the population may be excluded and others included. The list of subscribers

[3] From *Fifteenth Census of the United States, 1930, Instructions to Enumerators,* pp. 52-53.

for gas and electricity does not include the poorest homes in a city and will, therefore, not be an adequate basis to use for selecting a sample if we are studying unemployment or, in fact, any other topic which requires proper representation of the economic levels of the population.

In economic and social studies it is difficult to apply the mechanical methods necessary to obtain a random sample. Furthermore, our problem is complicated in that the units (persons, households, etc.) are dissimilar. When selecting marbles from a container, we do not care *which* white, black, or red marble we draw. We have units that differ from one another only in respect to color; they are made of the same material, are essentially the same size, shape, and weight, and have similar surfaces. When our units are people, we find that they differ in respect to sex, age, race, occupation, employment status, economic status, religion, etc. About all that they have in common is that they are human beings and live in the same community. Such differences are important and need to be kept in mind when a sample is selected. What has just been said should not be construed as a condemnation of the random sample; rather it is an attempt to point out the difficulty[4] of obtaining a random sample in particular instances, primarily when making a community study.

2. *Stratified sample.* A stratified sample differs from a random sample in that the population is broken into subgroups or strata before the sample is drawn. A random sample is then taken from each stratum. Usually the size of the sample from each stratum is proportional to the size of the stratum in the population. When a population is composed of relatively homogeneous units (such as the marbles referred to before), a random sample may be quite satisfactory. However, it frequently happens that a population is composed of heterogeneous units which, nevertheless, may be broken into rather uniform strata. The purchaser of a box of berries recognizes the existence of stratification when she turns out the contents to examine the bottom as well as the top layers. Here only two strata are considered. The purchaser of large quantities of coal will be apt to check his purchase in respect to lump size, heat content, ash content, etc. He is not satisfied with taking a few shovels full of coal from the top of a carload. The coal was probably loaded without any attempt to put small pieces on the bottom, but the shaking of the car on its journey from the mine tends to cause the smaller pieces to find their way to the bottom. Even though such a readjustment had not taken place, the careful buyer

---

[4] Sometimes Tippett's random numbers may be useful in selecting a sample if numbers can be assigned to the items in the population. See L. H. C. Tippett, *The Methods of Statistics* (2nd edition), p. 68, Williams and Norgate, London, 1937; and *Tracts for Computers*, XV, "Random Sampling Numbers," by L. H. C. Tippett, Cambridge University Press, 1927.

would select his sample from the middle and the ends and at various levels from top to bottom of the load in order to be sure of getting a sample as nearly as possible representative of the entire load.

The recognition of the existence of strata and the selection of random samples from these strata (rather than from the population as a whole) introduce added elements of control into the selection of the sample and give us greater assurance of representativeness, which increases as the number of strata increase. It will be shown in connection with Chapter XII that a stratified sample is more reliable than a random sample of the same size from the same population. From this it follows that the same reliability may be had from a smaller stratified sample. There is some danger that investigators, having an excessive feeling of security in the stratified sample, may depend too greatly upon the magic of stratification and use samples which are too small to give statistically reliable results. This can be guarded against by an intelligent use of the method and of the reliability formula given in Chapter XII. An extremely important point, which is often overlooked, is that the strata must be ones which are related to the topic being studied. If we are making a health study of male students in a college, we might recognize such strata as those who do or do not live at home; those who are totally, partially, or not at all self-supporting; those who do or do not take regular exercise; those who do or do not smoke; etc. However, there are other strata which clearly have no bearing on the problem. To take an extreme illustration, we might recognize such strata as those who habitually wear caps or hats, those who prefer single or double breasted coats, or any other categories which are not related to health.

The principle of the stratified sample is used by the American Institute of Public Opinion[5] in its surveys conducted throughout the country. The Institute refers to the method in its news releases as "scientific sampling." Seeking to measure public opinion not only in respect to elections but also on public questions of many sorts, the Institute at first used mail ballots, which were supplemented by direct interviews. Later the mail ballots were discontinued, and all opinions are now obtained through personal interviews. The Institute uses more than 600 field men in cities and rural areas throughout the nation. Voters are interviewed in the home, on the street, in offices, and on farms. To insure a representative sample, the Institute undertakes to select from the various strata a representative cross-section of the voters. Thus the sample (which appears

---

[5] This brief description of the procedure used by the American Institute of Public Opinion is based largely on a booklet issued by the Institute and entitled *The New Science of Public Opinion Measurement.*

to consist of more than 60 strata) must contain the proper proportion of:

 (1) Voters from each state (48 strata).
 (2) Men and women (2 strata).
 (3) Farm voters, voters in' villages of 2,500 population or less, and voters in urban communities divided into four categories according to population (6 strata).
 (4) Voters of all age groups (presumably several strata).
 (5) Voters of above-average and below-average incomes, as well as persons on relief (3 strata)
 (6) Democrats, Republicans, and members of other political parties, as indicated by how each person voted at the preceding presidential election (at least 3 strata).

The character of the cross-section of the sample is felt to be of more importance than the number of persons included. Straw votes and other sampling studies which are substantially wrong are usually incorrect rather because the persons reached are not representative of the entire group from which they were drawn than because the sample was too small. The failure of the *Literary Digest's* attempt to forecast correctly the 1936 election was due to the fact that its more than 2,300,000 ballots were not a representative cross-section. The voters were drawn primarily from lists of automobile owners and telephone subscribers. The Institute uses a sample of 3,000 to 50,000 or more cases, depending upon the problem being studied. When reporting the attitude of voters (or occasionally of some other special group) on a public question, the Institute does not ordinarily state the number of ballots obtained on that particular question. Presumably this practice is followed because the ordinary newspaper reader would think that a sample of a few thousand could hardly be depended upon to gauge the attitude of a nation, and he would be right if it were not for the careful application of the principle of the stratified sample.

The surveys conducted by the magazine *Fortune* also make use of stratified samples.[6] These studies are samples "balanced by geography, by sex, by size of community, by income group, by color, by age, and by occupation" in order to arrive at a true cross-section. This list includes two classifications (color and occupation) not used by the Institute, while the Institute considers how each person voted at the preceding presidential election, which is not included by *Fortune*.

Instead of selecting a proportionally stratified sample, it is sometimes easier to utilize a sample obtained by some other method and to adjust it so that each stratum will be properly represented. If all strata are represented, but not in the same proportions as in the population, weights may

---

[6] "The Fortune Quarterly Survey," *Fortune*, July 1938.

be applied to the portions of the sample coming from each stratum, in order to establish these proportions. Such a procedure, however, would not usually be so satisfactory as taking a stratified sample; moreover, it would presuppose a knowledge of the strata in the population and their importance, the same as for a stratified sample. A similar application of weights may be used when a sample, intended to be proportionally stratified, does not prove to have the proper proportions from each stratum.

3. *Other types of samples.* Sometimes a sample is selected by design or, as it is often termed, the selection is *purposive.* When selecting such a sample, the investigator sets out to make his sample agree in one or more respects with the population. Of course, this procedure cannot be followed unless the characteristics of the population are known. In a study of a group of wage earners a sample may be picked so that the average weekly earnings of those included in the sample will be the same as the average weekly earnings of the entire group of wage earners from which the sample was chosen. The sample might also be so chosen that it will agree with the larger group in respect to the average size of the wage earners' families. Additional controls could, of course, be used; if they are relevant to the problem which is being studied, the greater the number of respects in which the sample agrees with the population, the more thoroughly representative is the sample.

A stratified purposive sample may be employed if we first divide the population into strata and then endeavor to make the sample drawn from each stratum agree in one or more respects with all the items in that stratum, as well as to make the sample contain the proper number of items from each stratum.

Sometimes a sample is taken in a more or less haphazard fashion. Or, the investigator may include the data which are convenient or readily available, after which he will trustingly announce that the sample so taken is doubtless representative of the population which he is studying. For example, an investigator, who had ascertained that just under 2,500,000 children, eligible to be enrolled in high school, were not enrolled, desired to estimate how many of these 2,500,000 left school because of economic pressure. He managed to locate 16 acceptable studies concerning the reasons why students left school. These studies each included 53 to 274 children, a total of 2,525. The studies were made in schools in 13 different states. Negroes were studied in one instance. There were no figures from New York, Massachusetts, Illinois, Michigan, Wisconsin, Texas, and certain other populous states. Yet, because the geographical distribution was diverse and because large city, small city, and rural children were included, the investigator concluded: "The sample seems sufficiently representative of the various elements of the population to serve as the

basis for estimation of the whole group." This may or may not have been true. The sample was neither random, stratified, nor purposive; it merely included what was available.

As will be shown in connection with Chapter XII, the larger the sample (whether random or stratified), the more confidence we can place in conclusions drawn from the sample. It will also be shown that the greater the diversity there is in the population, the less reliability we can repose in samples of the same size. Mere size, of course, does not assure representativeness in a sample. A small random or stratified sample is apt to be much superior to a larger but badly selected sample. Sometimes a test of stability is made to determine when a sample is large enough. For example, a sample of 1000 may be selected from a group of voters, and 57.3 per cent may indicate they intend to vote for a certain candidate. Another 1000 may be chosen, and the two groups combined may show 56.9 per cent. Adding another 1000 may change the percentage to 56.8, and another 1000 (4000 in all) may leave the proportion unchanged, at 56.8. From this test, 3000 or 4000 would seem to be an adequate sample from the standpoint of size. However, the test of stability tests only stability and not representativeness. The fact that a percentage persists essentially unchanged means merely that we are continuing to get about the same result as before. Conceivably the first sample of 1000 could have been decidedly unrepresentative (say, from only the poorer sections of the voting population), and each succeeding sample similarly unrepresentative.

In selecting a sample it is important that bias be avoided. Bias does not mean the personal bias of the investigator which leads him to deliberately select his sample in order to show the results he desires. That is intellectual dishonesty. Neither does it mean that the persons answering the questions on the schedule are biased. The avoidance of bias involves, first, that there shall be no selective factor present in the drawing of the sample and, second, that there shall be no selective factor present when schedules are returned from those persons included in the sample. In the case of the *Literary Digest* 1936 straw vote, a selective factor was present because the basic lists from which the sample was selected did not include the lower economic levels of the population. Sometimes the basic list may be complete, but the method of selecting the sample may introduce bias. Thus a selection from an alphabetical list of names may be unsatisfactory because of nationality differences in the alphabetical distribution of family names. Such a bias may arise if sections of the list are chosen; it is not likely if (say) every tenth name is taken.

The second type of selective factor is frequently encountered if the questionnaire method of collection is used. When schedules are sent out by mail, an investigator never expects that all of them will be returned.

Since only part of the inquiries are answered, how can he be sure that those who did answer are representative of all those to whom schedules were sent? Often he cannot be sure; sometimes it is obvious that they are not representative. An alumni association sent out 363 inquiries to graduates, asking each to report (anonymously) his income for the preceding year. Replies· were received from 133. It is quite likely that a selective factor was present in these returns. Alumni who were out of work or who had very low incomes probably did not reply. This assumption is borne out by the data, which show an almost complete absence of incomes below $1,500, although the study was made in a depression year. Conclusions based upon biased samples are, obviously, not only useless but misleading.

Using the schedules to collect the data. When agents or enumerators take the schedules to the persons who are to furnish the information, the enumerators may explain the purpose of the investigation and solicit co-operation. Each question can be clearly explained as it is asked. Obviously, enumerators must be carefully instructed before they begin their work. Occasionally they are required to study the schedule and printed instructions, and then to take an examination. Enumerators should be persons of unquestioned integrity and should also be patient, polite, and tactful. Many a person resents being bothered to supply statistical (or other) information; some are reluctant; some refuse. The enumerator should plan his interviews to consume as little time as possible, and should bend every effort to get the desired information if it is feasible to do so. In some instances the work of the enumerator may be facilitated if a letter of explanation precedes the visit. Sometimes enumerators conduct interviews and fill in the schedules afterward. This is done on the theory that people feel more free to talk if the remarks are not being written down at the time. It is believed, however, that this is an undesirable procedure, especially when there are a number of facts to be remembered and later recorded. Enumerators should carry credentials in order that the persons visited may be satisfied as to the official connection of the visitor. Even though an enumerator makes his request for information as tactfully as possible, he may sometimes meet with a refusal. Frequently another visitor with a different approach may have better luck. It is sometimes a good plan to have one especially qualified worker who will follow up the more difficult cases.

Sometimes an enumerator may encounter a person who is too willing to cooperate and who wants to talk at great length about the study. In such a situation good terminal facilities are an asset. Carl Crow states[7]

---

[7] Carl Crow, *Four Hundred Million Customers*, pp 132–133, Harper and Brothers. New York. 1937.

that Chinese, when asked certain types of questions, are apt to give answers which they think will please the questioner. If an English investigating commission asks young Chinese where they want to go to school they are likely to reply, "England." The same author tells[8] of an investigation made in Amoy, where, because of a lack of proper death registration, the number of persons dying was estimated from figures of the number of coffins made. The figures of coffin production mounted, showing the development of an epidemic; but, after the epidemic was definitely known to have declined, the figures of coffins made remained high. Upon close inquiry it developed that the coffin manufacturers had continued to report peak production of coffins so that the agent of the health officials would not lose his job. They did not want to "break his rice bowl."

Sending schedules by mail rather than using enumerators is, at the outset, a less expensive method of collecting data. There is also the added advantage that the person supplying the information can fill out the form at his convenience, instead of being disturbed by the enumerator perhaps at a busy or inconvenient time. Furthermore, confidential information may be given in a questionnaire, which the informant would hesitate to divulge to an enumerator, provided of course that the informant is sure his identity is unknown. On the other hand, a large proportion of persons fail to reply to a mail inquiry (particularly certain classes of persons), and considerable follow-up work may be necessary. There is also great danger that the informant will not understand the questions, or will knowingly or otherwise make incorrect answers. Not only must clear, concise directions be sent with the schedule, but also a brief letter explaining the purpose of the inquiry and requesting cooperation. An addressed and stamped (or business reply) envelope should be included. An air mail business reply envelope (or card) is occasionally used by investigators with the hope that it will result in more and quicker responses. When follow-up work is necessary the persons who have not yet returned their forms may be sent courteous personal letters reminding them of the inquiry and again requesting cooperation. When appropriate, the follow-up may be by means of air mail letters, special delivery letters, registered letters (to be sure the communication has been delivered), telegrams, or telephone calls. Of course, the investigator should not make a nuisance of himself; he should not be too insistent When only part of the schedules are finally received, it is necessary to examine the situation carefully to be sure that no selective factor has been present. Or, if a selective factor appears to be present, it may be necessary to conduct a supplementary investigation to remedy the situation.

---

[8] *Ibid.*, pp. 252–253.

Editing the schedules. After the filled-out schedules are received, a certain amount of preparatory work is necessary before the data are in shape to be tabulated. The editorial tasks are varied. In the case of a small study one editor may do the entire work. In a larger study different phases of the editing may be portioned out among a number of editors.

1. *Computations.* It is usually better not to ask enumerators or persons supplying information to make any computations. Thus, if information has been obtained concerning the number of rooms in a home and the number of members in the household, the editor may compute the ratio of persons per room, to give some idea of crowding. If data have been collected concerning the time lost through non-compensated accidents and also of daily wages for each of a number of workers, the editor may compute for each case the income lost because of accidents.

2. *Coding.* Tabulation is frequently facilitated by coding. When machine tabulation (to be discussed shortly) is used, all entries on a schedule are reduced to a numerical code. Even when tabulation is manual, it may still be easier to look for a code mark—letters, numbers, or combination of letters and numbers—instead of attempting to read the original entry. The work of the tabulator may be further facilitated by the fact that the editor writes, or should write, legibly and uses a distinctive color, often red.

The Buffalo unemployment schedule on page 36 is shown edited according to a numerical code. Every entry is shown numerically coded (except those already expressed as numbers) in order to facilitate tabulation by mechanical means. The code scheme for industries (shown in column 6 of the schedule) ran as follows:

10. Professional
20. Clerical (not otherwise specified)
30. Domestic and personal service
40. Government employees (other than teachers)

Trade and Transportation

50. Retail and wholesale trade
51. Telephone and telegraph
52. Railway, express, gas, electric light
53. Water transportation
54. Bank and brokerage
55. Insurance and real estate
56. Other

Manufacturing and Mechanical Pursuits

60. Building trades, contractors
61. Building trades, wage earners

10497

Address 248 Blank Street

Territory 1-4

Visitor John Smith

Relation to head of household 1.	Sex 2.	Age 3.	Nativity 4.	Present or last regular employment — Employer 5.	Industry 6.	Occupation 7.	Employed now — Full time (√) 8.	Part time fraction 9.	Unemployed now — Weeks unemployed 10.	Reason for unemployment 11.	Able to work 12.	Wants to work 13.
a. Head **1**	M **1**	49	NW **1**	Standard Castings	64 Iron & steel	don't know	√ 10					
b. Son **2**	M **1**	26	NW **1**	Famous Dept Store dry goods	Retail **50**	salesmen	√ 10					
c. Daughter **2**	F **2**	24	NW **3**	Elite Barbers	**30** Tonsorial	manicurist	21 ½					
d. James **2**	F **2**	33	FB **3**	Famous Dept Store dry goods	Retail **50**	alterations	30		5	slack business	yes	yes
e.												
f.												
g.												

Notes

N. Y. State Dept. of Labor and The Buffalo Foundation   Unemployment Survey, Buffalo, N. Y. 1932

An Edited Schedule for a Hypothetical Family Based on the Code Used in the Buffalo Unemployment Study. Boldface numbers are editor's marks.

62. Clay, glass, and stone products
63. Food and kindred products
64. Iron, steel, and their products
65. Metal products, other than iron and steel
66. Paper, printing and publishing
67. Wearing apparel and textiles
68. Automobiles, parts, and tires
69. Lumber and furniture
70. Aeroplanes
71. Other manufacturing and mechanical pursuits

75. Labor (not otherwise specified)
80. Self-employed (other than 10 or 60)
90. Miscellaneous employments not classified above
00. Not reported

3. *Deciphering.* The handwriting of an enumerator or of an informant may occasionally be difficult to read. This is especially true when an enumerator makes entries on a schedule while he is outdoors in the rain or snow. Deciphering such copy is the editor's task; he not only saves time for the tabulator, but also insures accurate results. If entries are literally unreadable, the schedule may have to be referred back to the enumerator or the person who sent in the information.

4. *Checking entries.* The editor may look over the schedules for inconsistencies. Entries of age and date of birth may disagree. Something is probably awry if an individual reported as aged 8 is also shown to be married. Similarly, a mistake has probably (though not necessarily) been made if a woman is reported working full time as a blacksmith. Such entries must be verified if they are to be used.

5. *Examining for completeness.* The editor must also scrutinize the schedule to see if any entries are missing or incomplete. If the missing information is important, the schedule must be referred back to the enumerator or to the informant. Otherwise, the editor writes "N.R." (not reported) or a similar entry in place of the missing information.

Tabulating the data. After the schedules are edited, the data must be organized before finished tables and charts can be made. The following discussion treats of three methods that may be used.

1. *The score or tally sheet.* For purposes of illustration we shall assume that we want to show for the Buffalo study all males who were able and willing to work, classified by industry and by employment status. To simplify our illustration we shall consider employment status divided into three categories: employed full time, employed part time, and unemployed We shall not undertake, at this point, to subdivide part time employment.

or to classify unemployment according to duration or cause. The accompanying score sheet, or tally sheet, shows how this information could be assembled from a number of edited schedule cards. Note that it is more convenient and also saves space to use code numbers for the industries. As pointed out earlier, the numerical coding of the schedule card is shown because it is needed when mechanical tabulation is to be used.

AREA____1_____ SCORED BY __J. C. Williams__

DISTRICTS__1-5____ CHECKED BY __Paul Fry_____

INDUSTRY AND EMPLOYMENT STATUS, 1932
MALE, ABLE AND WILLING TO WORK

INDUSTRY	EMPLOYED FULL TIME		EMPLOYED PART TIME		UNEMPLOYED	
10	//// ////	④				
20						
30	// //	②	/	①	/	①
40	THL THL THL /// THL THL THL ///	⑱				
50	THL THL THL THL THL THL	⑮	// //	②	// //	②
51	/	①				
52	THL THL THL THL THL THL THL /// THL THL THL THL THL THL THL ///	㊳	////	④	///	③
53	/	①	/	①	/	①
54						
55						
56	THL THL // THL THL //	⑫	/	①	//	②
60	// //	②			/	①
61	THL THL THL THL	⑩	///	③	//// ////	④
62	// //	②	/	①		
63	THL THL	⑤	/	①	/	①
64	THL THL THL THL //// THL THL THL THL ////	24	THL THL THL THL THL THL THL THL	20	/// ///	④
65						
66	// //	②				
67						
68	THL THL	⑤			// //	②
69	/	①				
70	///	③				
71	THL THL	⑤			/	①
75					/	①
80	THL THL / THL THL /	⑪	/// ///	③	/	①
90						
00						

For hand tabulation (either by means of the tally sheet or by hand sorting, which is described in the following paragraph), we should probably code only the occupation and the reason for unemployment. Observe that the score marks are arranged in groups of five, four vertical and a diagonal. This facilitates counting. The data from the schedules are scored and then checked, and the totals of the tallies are entered. Since the tally sheet shown is for but one district, it is necessary to combine the results from a number of such tally sheets to arrive at the desired figures for the entire study. The resulting table is shown as Table 2.

TABLE 2

Employment Status of All Males Enumerated Who Were Able and Willing
to Work, Buffalo Unemployment Survey, 1932

Industry Group	Employed full time	Employed part time	Unemployed	Total
Professional .	197	18	19	234
Clerical (not otherwise specified)	1	1	36	38
Domestic and personal service	328	82	148	558
Government employees (other than teachers) . .	636	192	249	1,077
Trade and transportation .	1,843	734	883	3,460
Retail and wholesale trade	762	163	296	1,221
Telephone and telegraph .	34	20	24	78
Railway, express, gas, electric light	687	470	424	1,581
Water transportation	42	15	31	88
Bank and brokerage	99	6	20	125
Insurance and real estate. .	99	8	22	129
Other..	120	52	66	238
Manufacturing and mechanical pursuits	1,590	1,670	2,319	5,579
Building trades, contractors	87	115	177	379
Building trades, wage earners .	103	123	435	661
Clay, glass, and stone products .	17	27	37	81
Food and kindred products . .	338	100	131	569
Iron, steel, and their products	199	600	538	1,337
Metal products, other than iron and steel..	24	76	71	171
Paper, printing, and publishing .	117	69	50	236
Wearing apparel and textiles .	99	69	82	250
Automobiles, parts, and tires ..	212	224	435	871
Lumber and furniture....	79	65	119	263
Aeroplanes	95	13	69	177
Other	220	189	175	584
Labor (not otherwise specified) ...	4	11	16	31
Self-employed	653	86	122	861
Miscellaneous.	10	1	111	122
Total, males..	5,262	2,795	3,903	11,960

Source: Frederick E. Croxton, *Unemployment in Buffalo, November 1932*, p. 41, Special Bulletin No. 179, Division of Statistics and Information, New York State Department of Labor.

The tally sheet is a useful device for organizing information from a small study. It is apparent, however, that the tally sheet becomes cumbersome if there are many schedules to be scored or if it is desired to subdivide classifications. For example, if we desire to show how many men having part time work were employed less than $\frac{1}{4}$ time, $\frac{1}{4}$ but less than $\frac{1}{2}$ time, $\frac{1}{2}$ but less than $\frac{3}{4}$ time, and $\frac{3}{4}$ time but less than full time, it is necessary to divide the part time column of our tally sheet into five columns, four to accommodate the categories just mentioned and a fifth to take care of the "fraction not reported" group. If we wish to show duration of unemployment or cause of unemployment, we may subdivide the unemployed column. However, if we wish to show both duration and cause of unemployment so that duration may be studied in relation to cause, it is necessary to set up a new tally sheet, probably listing the causes vertically and the classified durations horizontally.

2. *Hand sorting.* When a study is fairly small and the schedule forms are small enough (and durable enough) to be handled readily, the data may be organized by a process of manual sorting. Hand sorting of the Buffalo cards is complicated by the fact that there are often several individuals on a card. This difficulty is overcome by organizing the data for heads of households first, then sorting for entries on line *b*, then for those on line *c*, etc. If we want to obtain the same sort of information as in the preceding paragraph, we might begin by sorting out the males and females. As a matter of fact, if tables are eventually to be made for females, it is more rapid to make an initial sort (for line *a*) into six piles: three for males—full time, part time, and unemployed; and three for females—full time, part time, and unemployed. The three piles of cards for males may then each be sorted into the industry categories. The cards in each pile are then counted to obtain the entries for a table similar to Table 2, but for male heads of households only. The next step would be to go through an identical sorting process for entries on line *b* of the schedule, then for line *c*, etc. The summation of all of these countings would result in Table 2. Details concerning part time employment or duration or cause of unemployment necessitate additional sorting.

3. *Mechanical tabulation.* Mechanical devices enable the work of tabulating a statistical study to be done most expeditiously, provided that the study is extensive enough to warrant their use. The use of tabulating equipment is recommended when there are a large number of schedules to be analyzed or there are numerous entries on each schedule. The process consists essentially of the following steps:

(a) Reducing all entries on the schedule to a numerical code.
(b) Recording these entries on a punch card by punching holes with a key punch to represent the code numbers.

(c) Sorting the cards by means of an electrical or mechanical[9] sorter.

(d) Assembling data from the sorted cards by means of a tabulator.

On page 42 is shown a punch card and also an enlarged portion of a card similar to the ones used for the Buffalo unemployment study. Each card was for an individual. The punch card shown here refers to the person listed on line *a* of the schedule shown previously. The first entry on the punch card (10437) covers five columns and refers to the number of the schedule so that the punch card and the schedule may be compared later if desired; furthermore the first digit, 1, tells us that this schedule came from area 1 of the study. There were nine areas in all. The next entry, using a single column, shows by a 1 that the individual was a head of a household (a 2 would indicate that he was not a head). The balance of the punch card is fairly obvious, except for the two columns marked "employment status." For these columns a two-digit code was devised which indicated whether the person was working full time or part time (and if so, what fraction), or was unemployed (and if so, the reason). There is no code number needed for column 11 of the schedule. Since the individual referred to on this punch card was working full time, the employment status columns are punched 10. Observe that it was necessary to use only part of the punch card for this study.

After the punch cards have been prepared they are sorted by the sorting machine (shown on page 43). If we wish to separate the cards for males and females, we set the machine to sort according to the numbers punched in column 7 of the punch card. As each card passes through the machine, electrical contact is made through the hole and the card is thus routed to the compartment which corresponds to its punched number. If the sort is for sex, we have cards in compartments 1 and 2 only. If we are sorting for age, we first sort for the right-hand digit, thus putting all ages ending in 0 in one compartment, all ending in 1 in the next compartment, and so on. The cards are then sorted according to the ten's digit. This sorting results in putting 19, 29, 39, 49, 59, etc., at the bottom of each compartment, after which 18, 28, 38, 48, 58, etc., drop on top of them, and so on.

When the cards have been sorted according to the categories we desire, the tabulation is completed by means of the tabulator (refer to page 43). This machine will not only count but also total several items at once and then print the results. Suppose we have sorted out all cards for unemployed people who were able and willing to work, and have arranged these by sex and by duration of unemployment. The tabulator will not only

---

[9] The devices shown and described here may be leased from the International Business Machines Corporation, 590 Madison Ave., New York City. Similar machines are available from Remington Rand Business Service, Inc., 315 4th Ave., New York City.

A Punch Card. These cards are 7⅜ x 3¼ inches in size.

A Portion of a Punch Card, Showing How the Data of the Buffalo Schedule on p. 36 Would be Recorded.
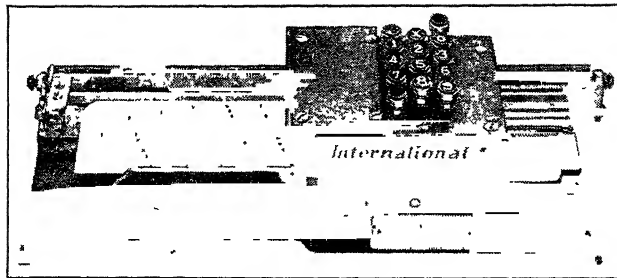
The Key-Punch.
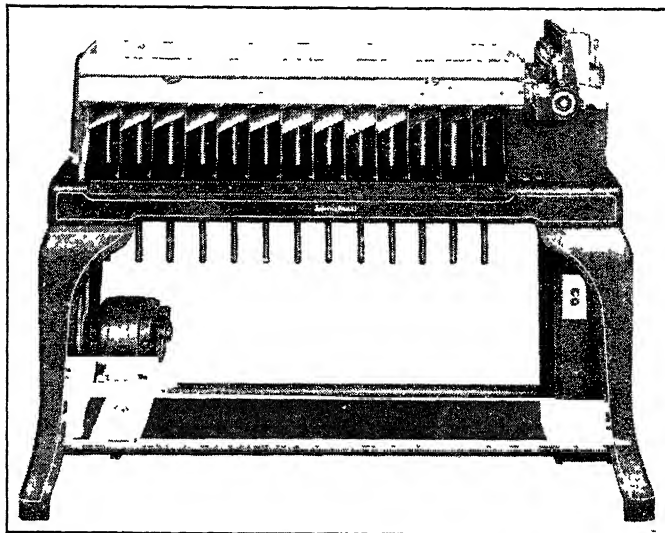
The Electric Sorter.   Cards are placed in the machine at the right and are then sorted into the compartments shown.
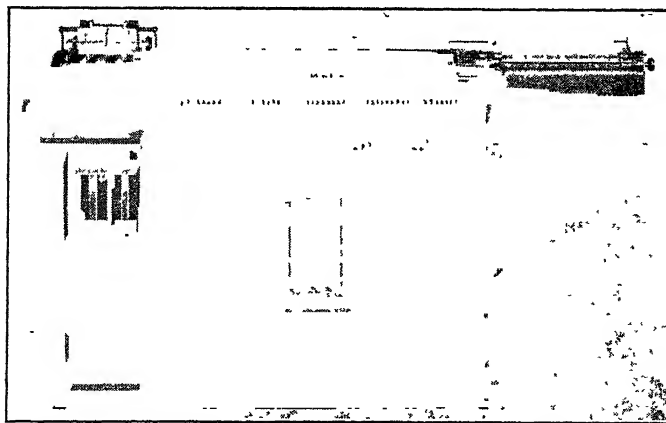
The Electric Printing Tabulator.   Cards are placed in the machine at the left; results are printed on paper at the right.

count the number of each sex in each duration category but will also accumulate the time lost, giving totals as needed. As a matter of fact the tabulator will prepare more detailed tables than those described above, but this description will suffice to give an idea of its operation.

A somewhat similar but much simpler device is sometimes used for small studies. It is known as the Findex[10] and consists of cards in which numbered holes are already punched. Facts are recorded by punching a slot to connect two adjacent holes, thus:

These cards are filed in a special cabinet, the front of which contains holes which correspond to the holes in the cards. Suppose that each card represents a workman, and that punching as shown above to connect the two holes means that this individual had an accident which involved a broken collar bone. Now, if we want to know how many such accidents occurred, we insert a rod through the upper hole, turn the case upside down (it is pivoted), and all cards recording broken collar bones will drop down about $\frac{1}{2}$ inch. The cards are then locked in that position, the case is righted, and the cards are counted. If we want to know how many cases of broken collar bones occurred to skilled workmen, we put a second rod through the appropriate hole so that, when the case is inverted, not all cards recording broken collar bones will drop but only those also referring to skilled workers.

Presentation and analysis. After the data have been organized manually or mechanically, the finished statistical tables and charts may be drawn up. Statistical tables are discussed in Chapter III, and charts in Chapters IV through VI. The analysis of data is treated in the remaining chapters of the book.

Statistical Sources

As pointed out at the beginning of this chapter, statistical data may already exist which will serve the purpose of the investigator. The data may or may not have been published. They may have been collected by an individual, a business firm, a research organization, a trade association, a newspaper or magazine, a government bureau, etc. Some organizations, such as the United States Census, issue only data which they themselves have collected. Such sources are designated as *primary*. Some publications bring together data originally compiled by others and are referred to as *secondary* sources. The *Survey of Current Business*, published monthly by the United States Bureau of Foreign and Domestic Commerce, is a secondary source as it includes data from numerous governmental and non-governmental sources. Obviously it is preferable to make

---

[10] Available from the Findex Co., Milwaukee, Wisconsin.

use of a primary source whenever possible, but it may often be more convenient to make use of a secondary source. One invaluable secondary source of data is the *Statistical Abstract of the United States,* issued annually by the Bureau of Foreign and Domestic Commerce. A number of other sources which are available in many libraries are listed in Appendix A, on pages 825–828.

The reasons for preferring a primary source are:

(1) The secondary source may contain mistakes due to errors in transscription when the figures were copied from the primary source.

(2) The primary source frequently includes definitions of terms and units used. This is an important consideration since intelligent use can hardly be made of data unless the user knows exactly what is meant by each term or unit employed by the collecting agency. When data are taken from several sources, it is particularly important that definitions of terms and units be scrutinized. The term "family" may sometimes have, the limited meaning of father, mother, and offspring; sometimes it may be used more or less synonymously with "household." The term "exports" may sometimes refer to gross exports (including re-exports); sometimes, to exports of United States' merchandise only. Although a measured bushel is 2,150.4 cubic inches, a bushel by weight does not represent the same number of pounds for all commodities. For example, a bushel of green peanuts in the shell weighs 22 pounds, a bushel of oats weighs 32 pounds, and a bushel of apples weighs 45 pounds; but a bushel of wheat, beans, peas, or potatoes weighs 60 pounds. The *Statistical Abstract of the United States,* although a secondary source, includes the necessary definitions of units.

(3) The primary source usually includes a copy of the schedule and a description of the procedure used in selecting the sample and in collecting the data; the reader is thus enabled to ascertain how much confidence to repose in the findings of the study.

(4) A primary source usually shows greater detail. A secondary source often omits part of the information or combines categories, such as showing counties instead of townships, or states instead of counties.

Reliability of data. The analyst should not make use of data, from either a primary or a secondary source, without assuring himself as to the reliability, accuracy, and applicability of the data. There are numerous points worthy of consideration here:

(1) If the enumeration was based on a sample, was the sample representative? Occasionally an investigator may select a sample of households at random from the lists of subscribers of a utility company, forgetting that the very poor may often not be users of gas or electricity and sometimes may not even possess a piped water supply.

(2) Was the schedule well designed? Were any leading questions or ambiguous questions included?

(3) Was the collecting agency unbiased or did it "have an axe to grind"? It is well to remember that bias may enter either consciously or unconsciously.

(4) Was a selective factor introduced because of careless enumeration? For example, in an unemployment study, canvassers might be careless about following up their calls at houses where no one was at home, and thus perhaps the data would show a smaller number of employed persons than actually existed.

(5) Were the enumerators capable and properly trained? Incompetent or poorly trained enumerators cannot be depended upon to produce useful results.

(6) Was the editing carefully and conscientiously done? Careless coding or computing, on the part of editors, may render of little value the findings of an otherwise valuable study.

(7) Was the tabulating (tally sheets, sorting, or mechanical tabulations) performed with care and accurately checked?

(8) In view of the definitions used, the area studied, and the methods of procedure, are the data applicable to the problem that is under investigation?

It is not always possible to ascertain the quality of work which was done by enumerators, editors, and tabulators. Most primary sources, however, reproduce a copy of the schedule used and generally give a more or less adequate description of the methods and procedures followed. Additional information may frequently be had by correspondence.

When using data over a period of years from a given source, we must be sure that definitions of terms have not changed or, if they have changed, to make due allowance for the change if it is possible to do so. For example: at the Census of 1910 and 1920, an *urban area* was defined as an incorporated place (including all towns [townships] in Massachusetts, Rhode Island, and New Hampshire) having 2,500 or more inhabitants. In 1930 the definition was modified first "to include townships and other political subdivisions (not incorporated as municipalities, nor containing any areas so incorporated) which had a total population of 10,000 or more, and a population density of 1,000 or more per square mile." This change affected 28 places in 5 states. A second modification, affecting New Hampshire, Massachusetts, and Rhode Island, included as urban the regularly incorporated cities and "only those towns in which there is a village or thickly settled area having more than 2,500 inhabitants and comprising, either by itself or when combined with other villages within the same town, more than 50 per cent of the total population of the town."

Comparability of different sources. When data are to be drawn from two or more sources, the reliability of each source must be considered and, in addition, the user must be sure that the data from the different sources are comparable. Let us list some of the reasons for lack of comparability:

(1) Different definitions of terms may have been used. Coal production is given by the United States Bureau of Mines in *short* tons of 2,000 pounds, while exports of coal are shown by the Bureau of Foreign and Domestic Commerce in *long* tons of 2,240 pounds. As if these two sorts of tons were not sufficiently confusing, it is necessary to be aware of two other "tons" used in shipping. These are the *gross* ton and the *net* (or registered) ton, each of which represents 100 cubic feet. Gross tonnage is the capacity of the hull plus the enclosed spaces above deck available for cargo, stores, passengers, and crew; whereas net tonnage is the gross tonnage less the space occupied by propelling machinery, fuel, crew quarters, master's cabin, and navigation spaces—in other words, approximately the space available for cargo and passengers.

Because of different accounting systems, the term "profit" may have different meanings in different industries. Profit for a railroad may be quite different from profit for a department store. In a certain industry, carried on almost solely by partnerships, an investigator found that many firms showed little or no profit and that great differences were present among firms. The partners were frequently paying themselves generous salaries, and therefore a new term "profit plus partners' salaries" was used for the study! Ages may be reported as of the last birthday; as of the nearest birthday; or, in Oriental fashion, as of the next birthday. Comparability of age data is thus affected by the bases of reporting.

(2) Different methods of computation or estimation may have been employed. For example, the methods of estimating population were responsible for two different estimates of the July 1, 1935, population of Yonkers, N. Y. One organization announced the population to be 144,233, while another estimated it as 157,455. The lower estimate assumed that Yonkers had grown since 1930 at the same rate as had the United States, the growth of the United States being determined by considering the excess of births over deaths and figures of net immigration. The second estimate appears to have been arrived at by assuming that the percentage change in the population of Yonkers from 1930 to 1935 was about one-half of the percentage change from 1920 to 1930.

(3) The samples may have been so chosen that the results are not comparable. Or, perchance, one study may have been based on a sample, whereas the other was a complete enumeration. It is, of course, possible to so choose a sample that the results of a study may be forced to fit a preconceived idea.

(4) Different standards of accuracy may have prevailed with respect to enumeration, editing, and tabulating.

(5) The sources may not be comparable in respect to areas included, or in respect to the period of time to which they refer. When the chronological difference is not too great, comparisons may sometimes be made or adjustments effected.

Whether an investigator is using primary or secondary sources, it is necessary to keep on the lookout for obvious mistakes and misprints. On page 392 of the *Statistical Abstract of the United States* for 1931, it is shown that in Continental United States potential water power amounting to 38,110,000 horse power is available 90 per cent of the time, while potential water power of 9,166,000 horse power is available 50 per cent of the time. It is clear that there must be a greater potential horse power available for 50 per cent of the time than for 90 per cent of the time. Data are given for each state and, if these details are added, it appears that 59,166,000 horse power of potential water power are available 50 per cent of the time. Obviously this was a typographical mistake which occurred in printing the abstract, or possibly was carried over from the primary source. Such an apparent contradiction would be observed at once by the experienced user of figures.

Selected References

W. B. Bailey and J. Cummings· *Statistics*, Chapters I–IV; A. C. McClurg and Co., Chicago, 1917. Chapter IV deals with editing.

F. E. Croxton and D. J. Cowden. *Practical Business Statistics*, Chapter II; Prentice-Hall, Inc., New York, 1934.

M. C. Elmer. *Social Research*, Chapters XVI, XVII, XX; Prentice-Hall, Inc., New York, 1939. Discussion of the sample, the interview, and the schedule.

G. A. Lundberg: *Social Research*, Chapters VI, VII; Longmans, Green and Co., New York, 1929. The schedule and the interview.

F. F. Stephan: "Practical Problems of Sampling Procedure," *American Sociological Review*, Vol. I, No. 4, August 1936, pages 569–580; and "Representative Sampling in Large-Scale Surveys," *Journal of the American Statistical Association*, Vol. 34, No. 206 (June 1939), pages 343–352.

P. V. Young: *Scientific Social Surveys and Research;* Chapters I–IV, VI–VIII; Prentice-Hall, Inc., New York, 1939. The social survey; the schedule and the interview.

CHAPTER III

STATISTICAL TABLES

Methods of Presentation

Four methods of statistical presentation are available. Data may be (1) incorporated in a paragraph of text, (2) put into tabular form, (3) placed in a semi-tabular arrangement, or (4) expressed graphically.

Text presentation. Combining figures and text is not a particularly effective device, since it is necessary to read, or at least scan, the entire paragraph before one can grasp the meaning of the entire set of figures. Most persons cannot easily comprehend the data when set forth in this manner, and it is especially difficult for the reader to single out individual figures. There is the advantage, however, that the writer can direct attention to, and thus emphasize, certain figures and can also call attention to comparisons of importance. Following is an example of text presentation:

> The United States Bureau of Foreign and Domestic Commerce presented, in the December 1937 *Monthly Summary of Foreign Commerce*, data of exports of United States merchandise and of imports for consumption (not including imports for purposes of re-export), segregated into "economic classes" and for various years. Comparing 1936 and 1937, the total value of exports v. - ‹2 ‵⁏‹ ‸‸‸ ‵‸‵‹⁏·ⁿ 1936 and $3,294,916,-000 in 1937, while the total value of imports for consumption was $2,423,977,000 in 1936 and $3,012,487,000 in 1937. Crude materials exported in 1936 amounted to $668,168,000, or 27.6 per cent of the total value of exports for that year; and in 1937 were $721,871,000, or 21.9 per cent of that year's total. Imports of crude materials amounted to $732,965,000 in 1936 and $973,535,000 in 1937, or respectively 30.2 per cent and 32.3 per cent of total imports for consumption in the two years. Crude foodstuffs exported in 1936 were valued at $58,144,000, which was 2.4 per cent of total exports for that year; and $101,742,000, or 3.1 per cent of the total, in 1937. Imports of crude foodstuffs for consumption were $348,682,000, or 14.4 per cent of the total value of imports for consumption in 1936; and $413,345,000, or 13.7 per cent of the total in

1937. Manufactured foodstuffs exported in 1936 came to $143,798,000, or 5.9 per cent of the year's total; and in 1937 were $177,451,000, or 5.4 per cent of the total. Imports of manufactured foodstuffs for consumption amounted to $386,240,000, or 15.9 per cent of the total imports in 1936; and $440,103,000, or 14.6 of the total in 1937. Semi-manufactures exported in 1936 were valued at $394,760,000, or 16 3 per cent of the total; in 1937 they were $677,254,000, or 20 6 per cent of the year's exports. Imports of semi-manufactures for consumption totaled $490,-238,000, or 20.2 per cent of all imports for consumption in 1936; and $634,181,000, or 21.1 per cent of the total in 1937. Finished manufactures worth $1,154,099,000, or 47.7 per cent of the total for that year, were exported in 1936; and $1,616,598,000 worth, or 49.1 per cent of the total, in 1937. Of finished manufactures imported for consumption $465,852,000 worth, or 19.2 per cent of all imports for consumption, came in during 1936 and $551,323,000, or 18.3 per cent of the total, were received in 1937.

Tabular presentation.   The same data that were included in the preceding text statement are shown in Table 3. This method of setting forth statistical data is usually superior to the use of text. A table with its title should be fully self-explanatory, although it may frequently be accompa-

TABLE 3

EXPORTS OF UNITED STATES MERCHANDISE AND IMPORTS FOR CONSUMPTION, BY ECONOMIC CLASSES, 1936 AND 1937

(Materials imported for purposes of re-export are not included)

Economic class	Value (thousands of dollars)		Per cent of total value	
	1937	1936	1937	1936
Exports of United States merchandise, total.	3,294,916	2,418,969	100 0	100 0
Crude materials.	721,871	668,168	21.9	27.6
Crude foodstuffs.	101,742	58,144	3.1	2.4
Manufactured foodstuffs.	177,451	143,798	5.4	5.9
Semi-manufactures.	677,254	394,760	20 6	16.3
Finished manufactures.	1,616,598	1,154,099	49 1	47.7
Imports for consumption, total..	3,012,487	2,423,977	100 0	100.0
Crude materials.	973,535	732,965	32.3	30.2
Crude foodstuffs	413,345	348,682	13.7	14.4
Manufactured foodstuffs	440,103	386,240	14 6	15 9
Semi-manufactures	634,181	490,238	21.1	20 2
Finished manufactures.	551,323	465,852	18.3	19.2

Source: United States Bureau of Foreign and Domestic Commerce, *Monthly Summary of Foreign Commerce*, December 1937. p. 36.

nied by a paragraph of interpretation or a paragraph directing attention to important figures.

It is readily seen that the table is much briefer than the text statement, since the row and column headings eliminate the necessity of repeating explanatory matter. As no text appears with the figures, the presentation is more concise. The logical arrangement of items in the stub (the left-hand column and its heading) and caption (the headings of the other columns) makes a table clear and easy to read. The use of columns and rows for the figures facilitates comparisons.

In Table 4 the various parts of the table have been slightly separated and labeled for identification. A table will have at least the four essen-

TABLE 4

POPULATION OF THE UNITED STATES, BY GEOGRAPHIC DIVISIONS, 1930 } Title

Division	Population	Per cent of total
United States.	122,775,046	100.0
New England.	8,166,341	6.7
Middle Atlantic	26,260,750	21.4
East North Central.	25,297,185	20.6
West North Central..	13,296,915	10.8
South Atlantic.	15,793,589	12.9
East South Central.	9,887,214	8.1
West South Central. .	12,176,830	9.9
Mountain...	3,701,789	3.0
Pacific	8,194,433	6.7

Caption labels the header row; Stub labels the Division column; Body labels the data.

Source note { Source· *Fifteenth Census of the United States, 1930*, Population Volume I, p. 10.

tials: title, stub, caption, and body. There may also be present a prefatory note (see Table 3) and one or more footnotes (see Table 7). If the figures in the table are not original, a source note is also included, sometimes with the prefatory note but usually below the table and below the footnotes to the table, if any are present.

Semi-tabular presentation. When only a few figures are to be used in a discussion, the text may be broken and the data listed as follows:[1]

. . . the employer "must gradually instruct his apprentice in the various operations of the trade to finally produce a competent worker." During the 2-year training period the apprentice must attend the courses in hygiene and related work at the university and obtain a certificate. The apprentice wage scale for a week's work is·

After 6 months at the school....	$ 7.50
After 12 months	10.00
After 18 months	12.00

[1] *Monthly Labor Review*, August 1935, p. 409. "Regulation of Beauty Shops under Quebec Labor Laws."

This method is not often used, but it is serviceable in that the figures are made to stand out from the text as they would not do if worked into one or two of the sentences. Incidentally, the figures can be more readily compared than if they were in the text.

Graphic presentation. Graphic devices are extremely useful and effective for quickly presenting a limited amount of information. The following chapters deal with curves, bar charts, maps, and other statistical diagrams.

Leading Considerations

Types of tables. From the point of view of usage there are two types of tables. In the first place there are *general* or *reference* tables, which are used as a repository of information. These are frequently very extensive, covering many pages, as, for example, Table 33 in Population Volume II of the 1930 Census, which takes up 18 pages. Such tables give detailed information arranged for ready reference. No attempt is made to arrange the entries in a general table so that emphasis will be placed on certain items, nor is there usually any reason for arranging columns and rows in order to bring out comparisons desired by the investigator. The primary, and usually sole, purpose of a reference table is to present the data in such a manner that individual items may be found readily by a reader. Reference or general tables are often placed in an appendix of a published report.[2]

In the second place there are *summary* or *text* tables, which are usually relatively small in size and which are designed to set forth one finding or a few closely related findings as effectively as possible. While the reference table may be rather complicated with subheadings and sub-subheadings in stub and caption, the summary table should be relatively simple in construction. It frequently accompanies a text discussion and hence is also referred to as a text table. If a reader is expected to divert his attention from a running discourse to a table, it is essential that it be not too formidable, but rather simple and easy to understand. Too many readers have a tendency to skip all the tables in a report, and this tendency can be combatted successfully only by making tables appear so simple as to be innocuous and by introducing simple and attractive graphs. Because of the purpose which a summary table is to serve, the items shown therein will be arranged to place emphasis where desired and the columns and rows will be so placed as to emphasize the comparisons of paramount importance.

---

[2] See, for example, Helen Herrmann, *Ten Years of Work Experience of Philadelphia Machinists*, Works Progress Administration, National Research Project and Industrial Research Department of the University of Pennsylvania, Philadelphia, 1938.

A summary table is almost invariably the result of boiling down information contained in one or more reference tables, although upon occasion a summary table may be based, in whole or in part, upon one or more other summary tables. Still more rarely a summary table may be constructed directly from data contained in schedule forms. The methods which can be used in deriving one table from one or more others are:

1. Data which are not important for the problem in hand may be omitted. Thus, although there are twenty-five states which produce bituminous coal, it might suffice to show separate data for only the leading ten or fifteen states.

2. Detailed data may be combined into groups. Thus data shown by states may be grouped into geographical divisions. Again, data shown by individual industries may be combined into broader industrial groups. For example, the manufacture of brick, tile, and terra cotta products; of cement, glass, and pottery; and the quarrying of marble, granite, slate, and like products may be combined into the major category "clay, stone, and glass products."

3. The arrangement of data may be altered. Thus an alphabetical arrangement of cities may be replaced by an arrangement according to size of municipality.

4. Averages, ratios, percentages, or other computed measures may be substituted for, or given in addition to, the original absolute figures. A column of ratios is shown in Table 8. It will be observed that these figures facilitate the interpretation of the data upon which they are based.

Comparisons. While the arrangement into columns and rows facilitates comparison of the data, such treatment does not automatically focus attention upon the comparisons that are important. This may be effected by placing the figures to be compared in contiguous columns or rows. Thus it may be seen that Table 5 facilitates the comparison of any one of the three items (number of returns, net income, or tax liability) in 1934 *with that same item* in 1935; whereas Table 6 makes it easy to compare number of returns, net income, and tax liability *with each other* for either 1934 or 1935. Either table enables us to compare one income class with another for a given year; however, such a comparison for two (or more) years is made more readily when the arrangement of columns is as given in Table 5.

Each of these tables is well constructed, but each focuses attention upon a different comparison. One of the most important considerations in table construction is that figures which are to be compared must be placed in immediate juxtaposition. It should be remembered that two or more series of figures are more easily compared when placed in adjacent columns than when placed in adjacent rows, and that figures of a series are more

TABLE 5

NUMBER OF INDIVIDUAL INCOME TAX RETURNS, AMOUNT OF NET INCOME, AND AMOUNT OF TAX, BY NET INCOME CLASSES, 1934 AND 1935

(Figures for net income and tax are in thousands of dollars)

Net income class	Number of returns		Net income		Tax	
	1935	1934	1935	1934	1935	1934
Total, all income classes .	4,575,012	4,094,420	14,909,812	12,796,802	657,439	511,400
Under $ 5,000	4,074,897	3,671,773	8,814,418	7,796,038	40,232	34,686
$ 5,000 and under 10,000	339,842	290,824	2,283,402	1,952,891	48,728	43,086
10,000 and under 25,000	123,564	102,892	1,822,271	1,513,592	103,754	83,960
25,000 and under 50,000	26,029	20,931	882,309	708,530	106,670	84,907
50,000 and under 100,000	8,033	6,093	535,772	405,976	112,816	84,792
100,000 and under 150,000	1,395	982	166,379	117,744	54,132	38,166
150,000 and under 300,000	896	690	179,911	140,960	74,039	57,995
300,000 and under 500,000	206	116	77,907	43,832	37,245	20,854
500,000 and under 1,000,000	109	86	73,811	59,464	38,323	30,745
1,000,000 and over	41	33	73,630	57,775	41,499	32,211

Source: United States Bureau of Internal Revenue, *Statistics of Income for 1936, Preliminary Report*, pp 8–10   This report contains data of 1936 returns filed through August 1937.

54

TABLE 6

NUMBER OF INDIVIDUAL INCOME TAX RETURNS, AMOUNT OF NET INCOME, AND AMOUNT OF TAX, BY NET INCOME CLASSES, 1934 AND 1935

(Figures for net income and tax are in thousands of dollars)

Net income class	1935			1934		
	Number of returns	Net income	Tax	Number of returns	Net income	Tax
Total, all income classes . . .	4,575,012	14,909,812	657,439	4,094,420	12,796,802	511,400
Under $ 5,000	4,074,897	8,814,418	40,232	3,671,773	7,796,038	34,686
$ 5,000 and under 10,000	339,842	2,283,402	48,728	290,824	1,952,891	43,086
10,000 and under 25,000	123,564	1,822,271	103,754	102,892	1,513,592	83,960
25,000 and under 50,000	26,029	882,309	106,670	20,931	708,530	84,907
50,000 and under 100,000	8,033	535,772	112,816	6,093	405,976	84,792
100,000 and under 150,000	1,395	166,379	54,132	982	117,744	38,166
150,000 and under 300,000	896	179,911	74,039	690	140,960	57,995
300,000 and under 500,000	206	77,907	37,245	116	43,832	20,854
500,000 and under 1,000,000	109	73,811	38,323	86	59,464	30,745
1,000,000 and over	41	73,630	41,499	33	57,775	32,211

easily compared with each other when arranged in a column than when placed in a row.

Comparisons may be greatly facilitated by the use of ratios, percentages, averages, or other computed relationships.   Ratios are shown in Tables 8 and 9; percentages, which are really a form of ratio (see Chapter VII), are included in Tables 3, 4, and 7.   Ratios and percentages are particu-

TABLE 7

POPULATION AND AREA OF CONTINENTAL UNITED STATES AND OUTLYING TERRITORIES
AND POSSESSIONS, 1930

Region	Population		Gross area in square miles
	Number	Per cent of total	
Total..	137,008,435	100.00	3,738,395
Continental United States	122,775,046	89.61	3,026,789
Outlying territories and possessions	14,233,389	10 39	711,606
Philippine Islands ..	12,082,366#	8.82	114,400
Porto Rico.	1,543,913	1.13	3,435
Hawaii Territory* . . . . . :..	368,336	.27	6,407
Alaska	59,278	.04	586,400
Panama Canal Zone	39,467	.03	549
Virgin Islands of the United States .	22,012	.02	133
Guam	18,509‡	01	206
American Samoa†	10,055	01	76
Military and naval, etc., services abroad	89,453	07	..

* Includes Midway Islands
† Includes Swain Island
# Estimated population July 1, 1929 (Thirteenth Annual Report of the Director of Education)
‡ Includes 1,118 persons on Naval Reservation, and on ships stationed at Guam.
Source: *Fifteenth Census of the United States, 1930*, Population Volume I, p 5

larly useful when the absolute figures to be compared are large.   Note in Tables 7 and 8 that rather large population figures can be compared readily by the use of percentages and ratios.   The use of averages is shown in Table 10; these averages would be particularly useful when compared with similar figures for other years.[3]   When tables show monthly fluctuations and both maxima and minima are noted, as in Table 10, the additional entry "Per cent variation of minimum from maximum" is useful to show the shrinkage from the high point to the low point during the year.

Emphasis.   The proper placing of an item in a table enables it to be given suitable emphasis.   Since occidentals read from left to right and from top to bottom, it follows that the most prominent position in the stub

---

[3] See, for example. *Monthly Labor Review*, January 1936 p. 49.

is at the top, and in the caption the most prominent position is at the left; likewise, the position of least prominence is at the bottom of the stub and at the right of the caption. Notice that, by following this principle in Table 3, exports were emphasized at the expense of imports, and 1937 was placed in a more prominent position than 1936.

TABLE 8

Population of the United States, by Sex and by Race and Nativity, 1930

Race and nativity	Total both sexes	Male	Female	Males per 100 females
Native white	95,497,800	48,010,145	47,487,655	101 1
Foreign-born white	13,366,407	7,153,709	6,212,698	115 1
Negro	11,891,143	5,855,669	6,035,474	97.0
Mexican	1,422,533	758,674	663,859	114.3
Indian	332,397	170,350	162,047	105.1
Japanese	138,834	81,771	57,063	143.3
Chinese . .	74,954	59,802	15,152	394 7
Filipino .	45,208	42,268	2,940	1437 7
Hindu	3,130	2,860	270	1059.3
Korean . . .	1,860	1,223	637	192 0
All other	780	609	171	356.1
Total	122,775,046	62,137,080	60,637,966	102.5

Source *Fifteenth Census of the United States, 1930,* Population Volume II, p 103

Totals are generally placed in either the most prominent or the least prominent position, depending upon whether or not it is desired to give emphasis to them. When "total" is shown at the top in the stub, a line should be placed below the first row of figures, as in Table 7. If the total entry is at the bottom of the stub, the figures are set off by a line drawn above them, as in Table 8. An alternate procedure consists of using a space instead of a line to set off the totals. Whatever its position, the word "total" in the stub should be indented if possible.

Individual figures, or columns or rows of figures, may also be emphasized by the use of boldface type, as in Table 9. When monthly fluctuations of employment, sales, or other factors are shown, the maximum figure may be set in boldface and the minimum may be put in italic type, as in Table 10. In general, italic is used to indicate an exception rather than for emphasis; thus italic type may be used for figures which are not to be included in taking a total.[4]

Arrangement of items in stub and caption. Considering the basic nature of statistical data which may be encountered, we have noted (p. 3)

---

[4] See Population Volume I, *Fifteenth Census of the United States, 1930,* p. 69.

that data may refer to geographical, chronological, qualitative, or quantitative classifications. We are now interested in the methods which may be employed in arranging the items in the stub or the caption (box head) of a table. The method of arrangement will be determined partly by the nature of the data (whether basically geographical, chronological, quali-

TABLE 9

Net Earnings and Changes in Total Capital Account of Insured Commercial Banks[1] during 1934

Earnings or change	Amount (in millions of dollars)	Amount per $100 of		
		Total available funds[2]	Total deposits	Total capital account
The banks' net earnings from current operations were	445	$1.03	$1 25	$ 7.27
Recoveries on assets written off and profits on securities sold were . .	290	.68	.82	4.73
Net earnings and recoveries were	735	1.71	2.07	12.00
The banks paid interest on capital notes and debentures, and dividends on preferred and common stock of .	185	.43	52	3.02
There remained after payment of interest and dividends	550	1.28	1.55	8.98
The banks wrote off losses of. .. .	1,080	2 52	3 04	17 64
The resulting reduction in capital account was	530	1.24	1.49	8.66
New capital funds were paid in to the net amount of	610	1.42	1.72	9.97
The net increase in the total capital account was[3].	80	.18	.23	1.31

[1] 14,124 banks; figures for 11 State banks in the District of Columbia, 2 insured national banks in Alaska and 9 other insured banks are not included Figures for national banks for second half of 1934 are estimated.
[2] Estimated average amount during year of total assets less customers' liability on account of acceptances, acceptances of other banks and bills sold with endorsement, and securities borrowed
[3] Exclusive of changes resulting from ! nening and closing of banks.
Source· *Annual Report of the Federal Deposit Insurance Corporation for the Year Ending December 31, 1934,* p 55.

tative, or quantitative), and partly by a consideration of whether the data are to appear in a reference table or in a summary table. A number of different methods of arrangement may be employed.

Alphabetical. This method of arrangement is admirably adapted for use in a general table, because it enables individual items to be located with ease. It is, obviously, not a useful method for text tables. It can be used only with series which are classified geographically or qualitatively.

Geographical. The geographical method of arrangement may be employed for series classified geographically, but it is applicable only when an established usage has been set up and should be used only when the statis-

TABLE 10

Number of Wage Earners of Both Sexes Reported Employed in Ohio Establishments on the 15th of Each Month, by Industry Group, 1937

(Reports are required from all mines and quarries and from all other concerns employing three or more persons. Both full time and part time employees are included)

Month	All industries	Agriculture	Construction	Manufactures	Mining and Quarrying	Service	Trade, retail and wholesale	Transportation and public utilities*
January........	997,641	7,524	35,089	678,023	33,062	119,489	65,950	58,504
February........	1,029,826	7,539	36,432	705,642	33,287	120,665	66,711	59,550
March........	1,051,082	8,382	38,673	719,137	33,629	122,332	68,504	60,425
April........	1,063,403	10,147	47,269	717,690	30,483	126,805	69,462	61,547
May........	1,095,245	10,728	53,260	738,546	30,677	129,352	69,857	62,825
June........	1,067,684	12,475	56,758	702,103	30,517	131,367	71,156	63,308
July........	1,093,490	13,482	60,014	726,013	30,287	129,166	71,140	63,388
August........	1,098,664	12,366	62,218	731,714	30,066	128,685	70,604	63,011
September........	1,109,800	13,042	61,242	735,868	32,294	131,722	72,148	63,484
October........	1,088,947	13,641	57,582	718,444	34,288	128,621	73,316	63,055
November........	1,016,682	10,290	50,733	663,800	34,093	125,968	70,624	61,174
December........	955,607	8,331	36,835	620,407	34,041	123,419	72,482	60,092
Average........	1,055,673	10,662	49,675	704,782	32,227	126,466	70,163	61,697
Per cent variation of minimum from maximum.	13 9	44 8	43 6	16 0	12 3	9.3	10 0	7 8

* Not including interstate transportation.
Source: Division of Labor Statistics, Department of Industrial Relations of Ohio.

tician is sure that his readers are familiar with the classification. In Table 4 the various geographic divisions of the United States are listed in their customary order. The various states in each division and their order of listing may be seen in Population Volume I, page 10, of the *Fifteenth Census of the United States, 1930.* Although the Census makes frequent use of the geographical method of arrangement for the states, it almost invariably lists the counties of a state alphabetically. For ease of reference, in a general table, the geographical arrangement is hardly so satisfactory as the alphabetical. While it may be argued that the geographical arrangement often places together contiguous, and therefore comparable, areas, it must be obvious that the geographical arrangement does not always do so. It is not usually a good method of arrangement for a summary table, since this arrangement does not place important items in prominent positions.

Magnitude. A very satisfactory method of arranging items in a summary table consists of listing them according to size, usually with the largest item first, but sometimes with the order reversed. The outlying territories and possessions shown in the stub of Table 7 are given in order of magnitude. When the largest item is placed first, the most important items (numerically) are placed in the most prominent positions. Arrangement of items according to size is not useful in a general table because it does not facilitate the finding of individual items as does the alphabetical arrangement. Data classified geographically or qualitatively may be arranged according to magnitude. So also may data classified chronologically, but they lose their chronological sequence when arranged by magnitude.

Historical. Data classified on a chronological basis would generally be arranged chronologically or historically. When years are listed, either the most recent or the earliest date may be shown first. The months, however, are usually listed with January first. When the historical arrangement is called for, it is adapted to either general or text tables. The historical arrangement is used in the stub of various tables in Chapter XV.

Customary. Certain data that are basically *qualitative* are generally arranged according to customary classes. The exports and imports of Table 3 are grouped into five categories: crude materials, crude foodstuffs, manufactured foodstuffs, semi-manufactures, and finished manufactures. In the stub of Table 8 the population of the United States is divided into ten customary groups upon a race and nativity basis, which are generally listed in the order shown, while in two of the box headings of this table the sexes are listed as male and female and are invariably given in this order. It will be noticed that following the listing of ten groups of the population there is given a category "all other." Such a group is usually placed at

the bottom in the stub, or at the right in the caption.   Good statistical practice dictates that an "all other," "miscellaneous," or "not reported" group should include relatively small numbers; otherwise the adequacy of the classification or the accuracy of the collection of the data may be questioned.   Arrangement by customary classes is appropriate for either a text or a reference table. *Quantitative* data may be arranged into classes as shown in the stub of Table 5.   Other arrangements of this type are shown in Chapter VIII.   Distributions of this sort generally begin with the class of smallest numerical value, as in Table 5.   Such an arrangement may be used in either a text or a reference table.

Progressive. This method of arrangement is illustrated in the stub of Table 9.   Notice that the items are listed in such a way that the final figure develops logically from those given before.   The table, as shown in the original report, contained no stub heading.   If items are related to such a degree that they can be placed together in a stub, it should ordinarily be possible to write a stub heading.   The student should consider what other headings might be appropriate for the stub of this table.   Another example of the progressive arrangement is given in the *Monthly Labor Review* for February 1939, page 357.   Monthly data are shown for the number of strikes, and the progressive headings in the caption are:

Con-tinued from preced-ing month	Begin-ning in month	In prog-ress during month	Ended in month	In effect at end of month

The progressive arrangement is suitable for either text or reference tables.

Numerical.   The wards of cities are usually designated as Ward 1, Ward 2, etc.   When data for such subdivisions are shown, a numerical arrangement is generally followed.   The precincts and districts of counties are sometimes numbered; the departments of a factory and salesmen's territories or sales areas may also be identified by numerical designations.   This method may appear in either a text or a reference table.   The numbers assigned to the categories are frequently only labels serving to identify some underlying arrangement.   For example, in a shoe factory, Department 1 was the cutting department; Department 2, the fitting department; Department 3, the lasting department, etc.

In using the various methods of arrangement, remember that the items should be arranged for greatest ease of reference in a reference table, whereas in a text table the arrangement should be designed to emphasize the important items and to stress the proper comparisons.

Details of Table Construction

Title and identification. A title should accompany every table and is customarily placed above the table. The title should be clearly worded and should state briefly what data are shown in the table. A title should be so worded as to mention the more important considerations first, placing toward the end any reference as to how the items are arranged and what period of time is covered. In general the title states, in order: what, where, how classified, and when. Illustrations of titles are shown in the various tables of this chapter. It will be noted that, when a title necessitates the use of several lines, an inverted pyramid arrangement is used.

If a title is unduly long, it may be advantageous to place a "catch title" above the main title or even to substitute the catch title for the full title. This shorter title undertakes merely to state the general nature of the data in the table. For Table 3 a catch title might read "FOREIGN TRADE, 1936 AND 1937."

When more than one table is included in a study, it is desirable to number the tables consecutively in order that each one may be identified by number rather than by title.

Prefatory note and footnotes. A prefatory note, one or more footnotes, and a source note may be appended to a table. A prefatory note is placed just below the title and in smaller or less prominent type. The prefatory note provides an explanation concerning the entire table or a substantial part of it, as in Table 10.

Explanations concerning individual figures, or a column or row of figures, should be given in footnotes. Footnotes keyed to stub entries and column headings may be referred to by means of numbers, as in Table 9; however, footnotes keyed to figures should be identified by a symbol (*, †, #, ‡, etc.), as in Table 7, or by a letter, but preferably not by a number.

Source notes. As previously indicated, the source note may appear below the title or below the footnotes. The latter practice has been generally followed in this text. The data set forth in a table will not often be material which the investigator has collected. Usually the figures will have been taken from one or more published or unpublished sources. The source note should be complete, giving author, title, volume, page, publisher and date. Not only is it courteous to mention the source of data quoted, but such information gives the reader some idea of the reliability of the data and makes it possible for him to refer to the original source to verify quoted figures or to obtain additional information. Although stating the source relieves the quoter of responsibility for shortcomings

in the original study, the practice does not relieve him of responsibility for knowingly or carelessly quoting unreliable data.

Sometimes data are taken from a secondary source instead of a primary source, because the secondary source may be more convenient. In such a case it may be advisable to mention both sources; for example, "Source: Automobile Manufacturers Association as quoted in *Statistical Abstract of the United States*, 1937, p. 363."

Data for a table may sometimes be taken from two or more different sources. When this is done, care must be exercised to see that the data are comparable. The importance of comparability of data was discussed in Chapter II; it is not necessary to say more on that topic at this point.

When apparent mistakes are found in a source, it is well to call attention to such difficulties. The December 1935 *Monthly Labor Review* (p. 1503) reprints a table from *The Oriental Economist* showing that total payrolls in 10 industries in Japan in 1933 were 647,340,199 yen, but points out in a footnote that, if the figures given for each of the 10 industries are added, the result is 647,430,199 yen.

Percentages. When percentages are used in a table, the stub or the caption entry should indicate clearly to what figures the percentages relate. Thus the term "per cent" alone should be avoided; rather say "per cent of total," "per cent of increase or decrease," etc. Sometimes tables are divided into a "number" section (showing amounts) and a "per cent" section, as in Table 3, which shows a "value" section and a "per cent" section. This table and Table 21 illustrate the use of adequate headings referring to percentages.

The percentages for the 1937 "exports" section of Table 3 total 100.1, while those for the 1936 "exports" section total 99.9. When individual percentages are written correct to tenths of one per cent, as is customary, the total will occasionally be slightly over or below 100.0 because of the accumulation of positive or negative remainders when rounding. If the percentages had been entered in hundredths or thousandths of a per cent, the total would have been closer to 100.0. Although a "per cent of total" column may add to slightly more or less than 100.0, the total is shown as 100.0, since that is what the individual percentages would yield if carried out far enough. If a total adds to less than 99.8 or more than 100.2, it is advisable to check the calculations for mistakes.

Rounding numbers. In order to avoid confusion and to facilitate comparisons, numbers of many digits may be rounded. Numbers may also be rounded because the compiler feels that they are accurate, not to the final digit, but only in terms of (say) thousands or millions.

Table 3 exhibits rounded numbers, and mention of that fact is made in

the caption. When numbers are rounded, a statement to that effect should be made in a prefatory note or in the stub or the caption. The wording may be "millions of . . . ," "000,000 omitted," and like expressions.

If a series of figures is to be expressed in thousands of dollars, for example, the rounding is to the *nearest* thousand. Thus $2,648,302 would become $2,648 (thousand) and $7,226,782 would become $7,227 (thousand). If the heading "thousands of dollars" appears in the box head (or stub) of a table, the dollar mark is not needed (see Table 3).

No serious error is ordinarily introduced by rounding. If each of a series of numbers is rounded, some will be raised and some will be lowered, but the errors so introduced tend to offset each other. Furthermore, it may be felt that to show all the digits of a large number is to give the appearance of spurious accuracy. For example, the population of the United States was ascertained to be 122,775,046 persons in 1930, but the figure could hardly be accurate to units or even to hundreds. However, it may be maintained that the figure 122,775,046 is the one obtained by the best methods available and is therefore probably more accurate than any rounded figure. Irrespective of the merits of these two points of view, six (or fewer) significant figures may often be accurate enough for the comparisons desired.

When computed values, such as totals, percentages, and averages, are to be shown in tables of rounded figures, these values should, if possible, be calculated from the original figures before rounding.

Totals. We have previously noted that totals, when of major importance, may be placed at the top in the stub and at the left in the caption. When it is not desired to emphasize totals, they may be placed at the bottom in the stub and at the right in the caption.

Table 8 carries both a total column and a total row. An arrangement such as this results in a single number (122,775,046) which is sometimes termed a "grand total" or a "checked grand total." The fact that the figures yield the same sum when added vertically and horizontally is not a positive check since two or more compensating errors may have been made. That, however, does not often happen. We do have definite proof either that no errors were made or that more than one was made.

Units. The units of measurement of the figures in a column or a row of a table may often be self-explanatory. When this is not true, the nature of the unit should be made clear in the stub or the box head, as in Table 9. If the explanation applies to all figures in the table, it may appear as a prefatory note. Data of monetary units are usually self-descriptive, because of the use of the dollar sign. Note, in the last three columns of Table 9, that this sign appears for only the first entry in a column.

Size and shape of table. In general a table should be designed so that

it will be neither very long and narrow nor very short and wide. A table must also be adjusted to the space in which it is to appear. Usually this limitation takes the form of a page of a book or a report. Of course, a table need not occupy the entire length or width of a page. If the table is too large for the allotted space, it may be recast into several smaller tables. Reduction of type size may permit a table to be included on a page, but reduction should not be made at the expense of legibility. If the use of a folded page is not desirable, the table may be arranged to occupy two facing pages. Because of the difficulty of aligning pages perfectly in binding, the stub should be repeated on the second page. When reference tables are continued over several pages, they may be split either vertically or horizontally. In either case complete stub and caption entries should appear on each page, the title should be repeated on each page, and footnotes may appear at the bottom of the appropriate page or may be accumulated at the end of the table.

The horizontal dimension of a table may be determined by allowing for:

(1) Width of stub, determined by longest entry. (A very long entry may be put on two or more lines to save space, see Table 9.)

(2) Width of each column, determined by largest number or by entry in each box head. (By hyphenating words an entry in a box head may be compressed horizontally and expanded vertically.)

(3) Ruling.

(4) Margins.

The vertical dimension may be ascertained by considering:

(1) Space needed for title, prefatory note, footnotes, and source note. Since the first line of the title should not exceed the table in width, a long title may require several lines.

(2) Number of lines needed for caption or stub heading.

(3) Number of rows in body of table.

(4) Ruling.

(5) Margins.

Ruling. Most of the tables in this text are shown with single-line ruling and are open at the sides. Double-line ruling is sometimes used but, to the writers at least, seems to make either hand-ruled or printed tables appear somewhat complicated. Tables are rarely closed at the sides, and should never appear with one side closed and one open.

There seems to be a growing tendency to use text tables without ruling, either vertical or horizontal. Table 11 shows how Table 8 appears when no ruling is used.

An examination of tables in this book and elsewhere will show that:

(1) No horizontal lines are used in the body of a table except to set off totals and occasionally to separate a table into distinct parts.

(2) Horizontal lines separating major and minor box heads do not continue into the stub heading.

TABLE 11

POPULATION OF THE UNITED STATES BY SEX AND BY RACE AND NATIVITY, 1930

Race and nativity	Total both sexes	Male	Female	Males per 100 females
Native white . . .	95,497,800	48,010,145	47,487,655	101.1
Foreign-born white.	13,366,407	7,153,709	6,212,698	115.1
Negro. . .	11,891,143	5,855,669	6,035,474	97 0
Mexican.	1,422,533	758,674	663,859	114.3
Indian	332,397	170,350	162,047	105.1
Japanese.	138,834	81,771	57,063	143.3
Chinese	74,954	59,802	15,152	394.7
Filipino.	45,208	42,268	2,940	1437.7
Hindu	3,130	2,860	270	1059.3
Korean	1,860	1,223	637	192.0
All other	780	609	171	356.1
Total......	122,775,046	62,137,080	60,637,966	102.5

Source· *Fifteenth Census of the United States, 1930*, Population Volume II, p. 103

(3) All vertical lines separating box heads appear only between the box heads which they separate; they do not extend above these box heads.

Guiding the eye. Skipping a line every three, four, or five rows, as in Table 10, makes it easier for the eye to follow the rows across a table. The use of leaders in the stub of a table is also helpful.

Zeros. It is not customary to show a zero in a table (other than a computation form). When no cases have been found to exist or when the value of an item is zero, the fact may be indicated by means of dots (. . .) or short dashes (—). When there is no figure for an entry because information is lacking, a footnote should be used to indicate that fact.

Size and style of type. Too much variety in size or style of type (or lettering) is not desirable. In general the title should be most prominent and is usually set in large and small capitals or in bold face type. The items listed in the stub and caption and the figures in the body of the table are usually set in the same size type. Footnotes, prefatory note, and source note are generally set in smaller type than that used in the body of the table.

Statistical Reports

When making a statistical report, the method of preparing the tables will be dictated partly by the number of copies of the report required and partly by the cost involved. Tables may be handwritten, typewritten,

mimeographed, multigraphed, reproduced by a photostatic or photographic process from handwritten or typed tables, or printed.

There is a distinct disadvantage in the use of the ordinary typewriter for preparing other than relatively simple tables, because of the lack of flexibility of spacing and of size of type. Table 12 shows a table without ruling, prepared on an ordinary typewriter with pica type. Table 13

TABLE 12

EXPORTS OF UNITED STATES MERCHANDISE AND IMPORTS FOR CONSUMPTION, BY ECONOMIC CLASSES, 1936 AND 1937

(Materials imported for purposes of re-export are not included.)

Economic class	Value (thousands of dollars)		Per cent of total value	
	1937	1936	1937	1936
Exports of United States merchandise, total	3,294,916	2,418,969	100.0	100.0
Crude materials	721,871	668,168	21.9	27.6
Crude foodstuffs	101,742	58,144	3.1	2.4
Manufactured foodstuffs	177,451	143,798	5.4	5.9
Semi-manufactures	677,254	394,760	20.6	16.3
Finished manufactures	1,616,598	1,154,099	49.1	47.7
Imports for consumption, total	3,012,487	2,423,977	100.0	100.0
Crude materials	973,535	732,965	32.3	30.2
Crude foodstuffs	413,345	348,682	13.7	14.4
Manufactured foodstuffs	440,103	386,240	14.6	15.9
Semi-manufactures	634,181	490,238	21.1	20.2
Finished manufactures	551,323	465,852	18.3	19.2

Source: United States Bureau of Foreign and Domestic Commerce, *Monthly Summary of Foreign Commerce*, December 1937, p. 36.

presents the same data and indicates how ruling may be done on a typewriter. Note that more flexibility was obtained by using two typewriters, one with pica and one with elite type. By using elite type for the stub entries and the body, a certain amount of space may be saved. Somewhat more flexibility in planning a table may be had by using a typewriter with variable spacing and with different kinds and sizes of type.

If only a few copies of a report are to be made and if the tables are simple, the tables and accompanying text may be typed and carbon copies

made.   If several dozen copies are needed, the longhand or typed material may be photostated at a cost of about 25 cents per $8\frac{1}{2} \times 11$ inch page.   By this method, reduction or enlarging is possible and copies may be had rather promptly since no plate need be made.   If a larger number of copies is required, resort may be had to mimeographing or multigraphing.

TABLE 13

EXPORTS OF UNITED STATES MERCHANDISE AND IMPORTS FOR CONSUMPTION, BY ECONOMIC CLASSES, 1936 AND 1937

(Materials imported for purposes of re-export are not included.)

Economic class	Value (thousands of dollars)		Per cent of total value	
	1937	1936	1937	1936
Exports of United States merchandise, total	3,294,916	2,418,969	100.0	100.0
Crude materials	721,871	668,168	21.9	27.6
Crude foodstuffs	101,742	58,144	3.1	2.4
Manufactured foodstuffs	177,451	143,798	5.4	5.9
Semi-manufactures	677,254	394,760	20.6	16.3
Finished manufactures	1,616,598	1,154,099	49.1	47.7
Imports for consumption, total..............	3,012,487	2,423,977	100.0	100.0
Crude materials	973,535	732,965	32.3	30.2
Crude foodstuffs	413,345	348,682	13.7	14.4
Manufactured foodstuffs	440,103	386,240	14.6	15.9
Semi-manufactures	634,181	490,238	21.1	20.2
Finished manufactures	551,323	465,852	18.3	19.2

Source: United States Bureau of Foreign and Domestic Commerce, Monthly Summary of Foreign Commerce, December 1937, p. 36.

Tables may also be reproduced by a photo-offset process, which is quite satisfactory and is often cheaper than printing because it avoids the type-setting.   Enlarging or reduction is possible; typed material may be reduced so that 4 ordinary $8\frac{1}{2} \times 11$ inch pages (pica type) will appear on one page.   It should be noted that the typed copy should be a first-class job if satisfactory reproductions are to be obtained.

Occasionally the gelatin-pan method may be useful when only a few copies are needed.   A special ink is available for handwritten material and for illustrations; also, ribbons and carbon paper may be obtained for typed material.   This method is hardly so satisfactory as those above mentioned, but it enables a few copies to be made by anyone with rather inexpensive equipment.   The method does not enlarge or reduce.

Selected References

F. E. Croxton and D. J. Cowden: *Practical Business Statistics,* Chapter III; Prentice-Hall, Inc., New York, 1934.

W. L. Crum, A. C Patton, and A. R. Tebbutt· *Introduction to Economic Statistics,* Chapters V, VI; McGraw-Hill Book Co , New York, 1938. General tables and summary tables are discussed separately.

B. D Mudgett· *Statistical Tables and Graphs,* Chapters I–III; Houghton Mifflin Co., Boston, 1930.

United States Department of Agriculture, Bureau of Agricultural Economics: *The Preparation of Statistical Tables, A Handbook,* Washington, 1937 A mimeographed set of recommendations designed to promote uniformity of tables prepared by the Bureau of Agricultural Economics.

H. M. Walker and W. N. Durost: *Statistical Tables, Their Structure and Use,* Teachers College, Columbia University, New York, 1936. Educational illustrations.

CHAPTER IV

GRAPHIC PRESENTATION
SIMPLE CURVES

The Graphic Method

Attention has already been given to the presentation of statistical data by means of text, tabular, and semi-tabular devices. Ordinarily statistical data will be presented either in the form of a table or a chart. This chapter and the following two are devoted to a discussion of the graphic devices by means of which statistical data may be set forth. As will be readily seen from a perusal of the pages of this book, charts or graphs are more effective in attracting attention than are any of the other devices. Readers are therefore not so likely to skip a chart as a table. A simple, attractive, well-constructed graph, showing a limited set of facts is easier to understand than is a table.

The outstanding effectiveness of a chart as a device for presenting a limited amount of data, makes it a most useful statistical tool. Certain limitations should be noted, however. In the first place, charts cannot show so many sets of facts as may be shown in a table. Numerous columns and rows may appear in a table; but imagine Chart 3 (or any other in this chapter) with six or eight criss-crossing and intertwining lines, and it is immediately obvious why a chart should show a limited amount of information at once. In the second place, although exact values can be given in a table, only approximate values can ordinarily be shown by a chart. In a table we may enter as many digits as desired, but we can plot only the approximate value on a chart. For example, while the data upon which Chart 3 is based could be recorded in a table in terms of dollars and cents, no such accuracy is possible in the chart. Thus charts are useful for giving a quick picture of the general situation but not of details. In the third place, charts require a certain amount of time to construct, since each one is an original drawing. This difficulty, however,

is offset by the added effectiveness which the chart possesses in comparison with a table.[1]

Types of Charts

In this text we shall discuss: *curves* or *line diagrams; bar charts,* involving one-dimensional comparisons; *area diagrams,* involving two–dimensional comparisons (including particularly *pie diagrams* which involve one- or two-dimensional comparisons, or comparisons of angles); *volume diagrams,* involving a visualization of the third dimension and three-dimensional
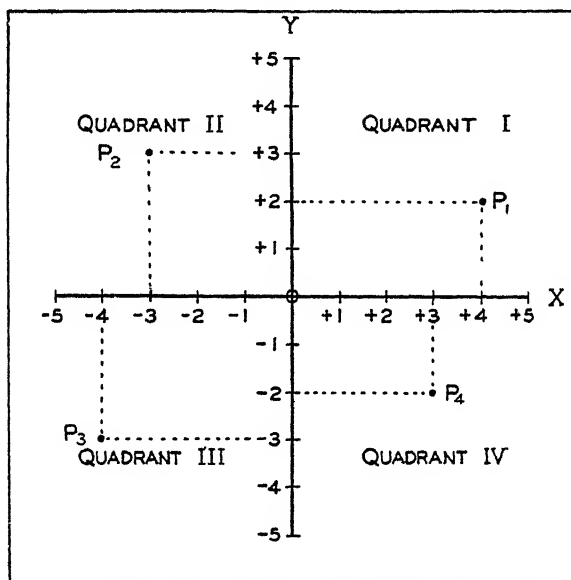
Chart 2.   Axes for Curve Plotting.

comparisons; and *statistical maps.* Other specialized types of charts and certain charts which are graphic but not statistical (for example, organization and procedure charts) are not treated here, but are discussed in some of the special references on graphic methods listed at the end of this and the following chapters.

[1] William Playfair, who is understood to have "invented outright" the graphic method in the latter part of the 18th century says: "The advantage proposed by this method, is not that of giving a more accurate statement than by figures, but it is to give a more simple and permanent idea of the gradual progress and comparative amounts, at different periods, by presenting to the eye a figure [chart], the proportions of which correspond with the amount of the sums intended to be expressed." See the article "Playfair and His Charts," by H. Gray Funkhauser and Helen M. Walker, in *Economic History,* February 1935, pp. 103–109.

Plotting a Curve

When statistical data are shown as curves, the points are plotted in reference to a pair of intersecting lines, called axes and shown in Chart 2. The horizontal line is known as the "X-axis," and the vertical line is designated as the "Y-axis." Positive values are shown to the right of zero on the X-axis and above the zero on the Y-axis; negative values are placed to the left of zero on the X-axis and below the zero on the Y-axis. The point at which the two axes intersect is zero for both X and Y and is referred to as the "zero point," the "point of origin," or merely the "origin." The positive and negative values on the axes increase as we move away from this origin.

MILLIONS
OF DOLLARS
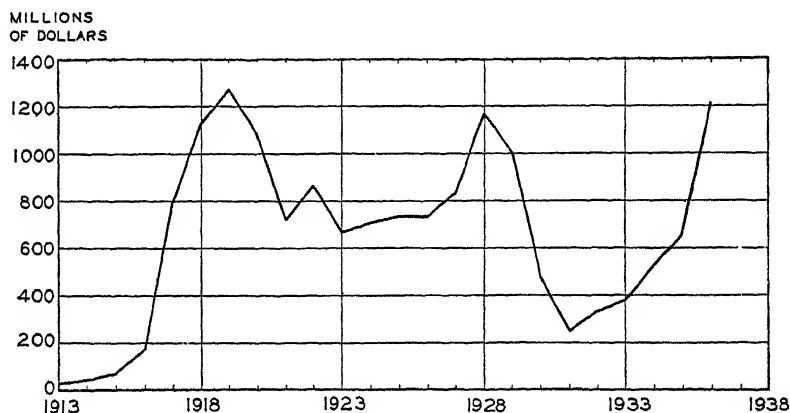
Chart 3. Individual Federal Income Taxes in the United States, 1913-1936. The reader may wish to add the later data to this chart as they become available. (Data from *Statistical Abstract of the United States, 1937*, p. 117, and by correspondence.)

When plotting two variables on the axes, one variable is assigned to one axis and the other to the other axis. It is customary to put the values of the independent variable on the X-axis and the values of the dependent variable on the Y-axis. In Chart 3, for example, it is apparent that time is the independent variable; therefore time appears on the X-axis. Income tax payments, on the other hand, constitute the dependent variable; hence they are shown on the Y-axis. In Chart 4 the independent variable is the amount of income received by dentists, while the dependent variable is the proportion of dentists receiving various specified incomes. In some cases two variables may be mutually dependent on each other, or they may both be dependent upon some other factor. Under such conditions one variable is more or less arbitrarily placed on one axis and the other variable on the other axis.

The two axes divide the plotting area into four sections known as "quadrants." For reference purposes these quadrants are designated I, II, III, and IV, as shown in Chart 2. Quadrant I accommodates values which are positive on both the X- and Y-axes. Quadrant II provides for values which are negative on the X-axis and positive on the Y-axis. Quadrant III takes care of values which are negative on both axes. Quadrant IV is for values which are positive on the X-axis and negative on the Y-axis.

Any point plotted in one of the quadrants may be located by referring to its abscissa, which is its horizontal or X distance from zero, and to its ordinate, which is its vertical or Y distance from zero. For illustrative
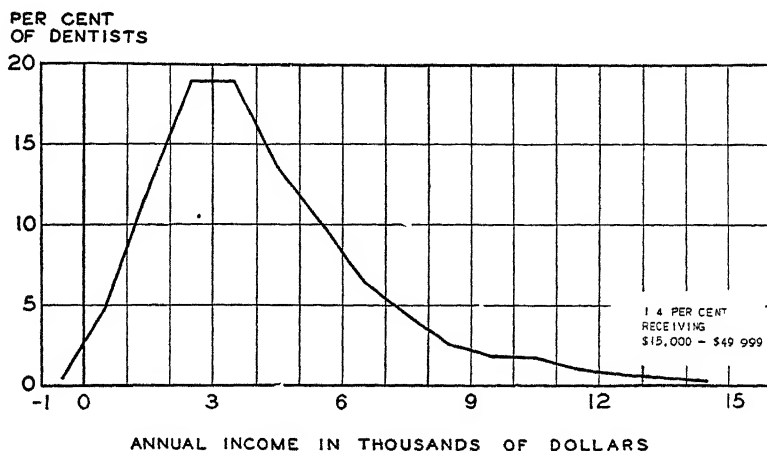
Chart 4. Annual Net Incomes of Dentists Engaged in General Practice, 1929. [Data from Maurice Leven, *The Incomes of Physicians*, p. 131. The University of Chicago Press, Chicago, 1932.)

purposes four points have been plotted on Chart 2, one in each quadrant. P_1 represents $X = +4$, $Y = +2$. P_2 indicates $X = -3$, $Y = +3$. P_3 is $X = -4$, $Y = -3$. P_4 shows $X = +3$, $Y = -2$.

When the axes are used as bases of reference for plotting equations, any or all of the quadrants may be used, since many equations may call for negative values of X or of Y, or of both. At present, however, we are not interested in the graphic representation of equations (see pp. 395–397 and 426–432), but in graphically portraying observed statistical data. When we are dealing with statistical data, it must be obvious that both the X and Y variables are ordinarily positive quantities, and that therefore we shall generally use only the quadrant designated as I. Chart 3, show-

ing the individual income tax payments in the United States over a period of years, is an example of a curve lying wholly in quadrant I.

Quadrants II and IV are occasionally used in conjunction with quadrant I. Chart 4 shows a curve which makes use of quadrants I and II; the curve of Chart 5 lies partly in quadrant I and partly in quadrant IV. Since both X and Y values are negative in quadrant III, that quadrant is very rarely used.

Types of Data Shown by Curves

It was noted earlier that statistical data are basically classified according to chronological, geographical, quantitative, or qualitative characteristics.
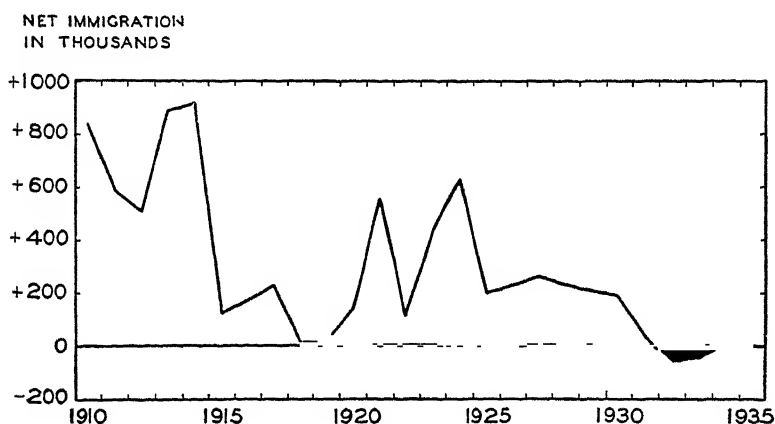
Chart 5. Net Immigration (Immigration Minus Emigration) into the United States, 1910-1935, Years Ending June 30. (Based on data of Chart 23.)

Curves are frequently used for picturing time series and for showing frequency distributions (by far the most important sort of quantitatively classified data), although, of course, other types of graphs are also applicable as shown in the following chapters. Qualitatively and, especially, geographically classified data are rarely depicted by curves; instead, bar charts and other devices are used, as will be indicated hereafter.

Time series curves. The method of plotting time series depends upon the type of data to be represented. We may distinguish between *period data* and *point data*. Period data, such as total sales per month, average monthly sales per year, and average prices during the year, refer to a period of time. Point data are those, such as inventory values, price quotations, or temperature readings, which refer to a particular point of time.

Charts 3 and 5, which represent period data, have dates along their horizontal scales, as is customary with all time series charts (whether they are of period data or point data).   When annual data of this type are plotted, the dates on the horizontal scales may be placed below the vertical lines, as in Chart 3, or below spaces as in Chart 5 and as in the left-hand part of Chart 29.   Either method may be used; one argument for labeling the spaces is that this gives a visual impression of time as having duration. When monthly (and daily, weekly, or quarterly) data are plotted over a period of years, there is no choice but to label the spaces representing each year since, if the lines were labeled, it would not be immediately obvious

DOLLARS
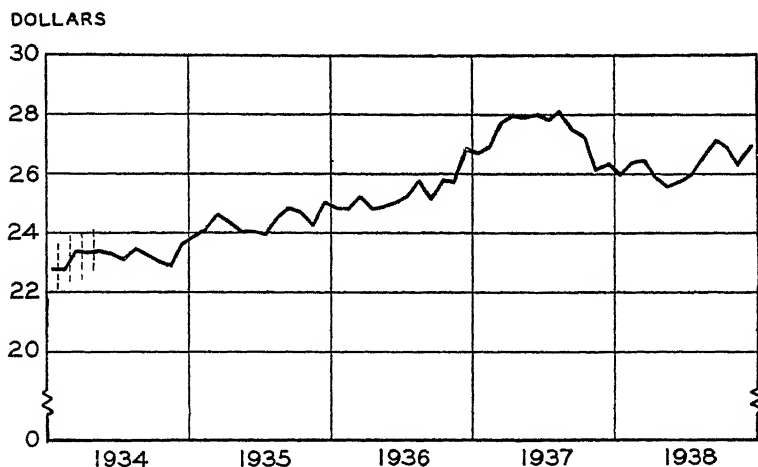
Chart 6.   Average Weekly Earnings of Employees in New York State Factories, 1934-1938.   (New York Department of Labor, *The Industrial Bulletin*, January 1939, p. 23)

to all readers whether the label referred to the space preceding the label. the space following the label, or possibly half of the space on each side. Each horizontal year space is divided into 12 parts for the plotting of the monthly figures, and these figures are plotted at the middle of each of the 12 spaces.   Chart 6 is an illustration of this procedure.

When point data are used, label the spaces and plot the observation within that space at the point of time to which the data refer.   Thus, in plotting monthly data, we should plot the figures at the beginning of each space representing the month for beginning-of-the-month data, at the middle of the space for middle-of-the-month data, and at the end of the space for end-of-the-month data.   If this scheme is followed, the results are as shown in Charts 7 and 8; Chart 7 shows first-of-the-month data and

MILLIONS
OF POUNDS
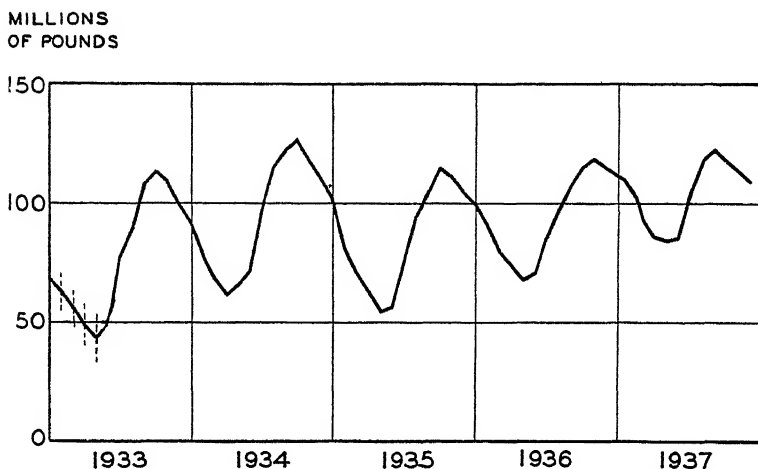
Chart 7.   Cold Storage Holdings of Cheese as of the First of Each Month, 1933–1937
(Data from *Agricultural Statistics*, 1938, p. 363.)

Chart 8 shows end-of-the-month data.   Point data as of the middle of the month, such as employment data which are often collected as of the payroll period nearest the fifteenth of each month, would be charted in similar fashion to the data in Chart 6, which shows period data.

DOLLARS
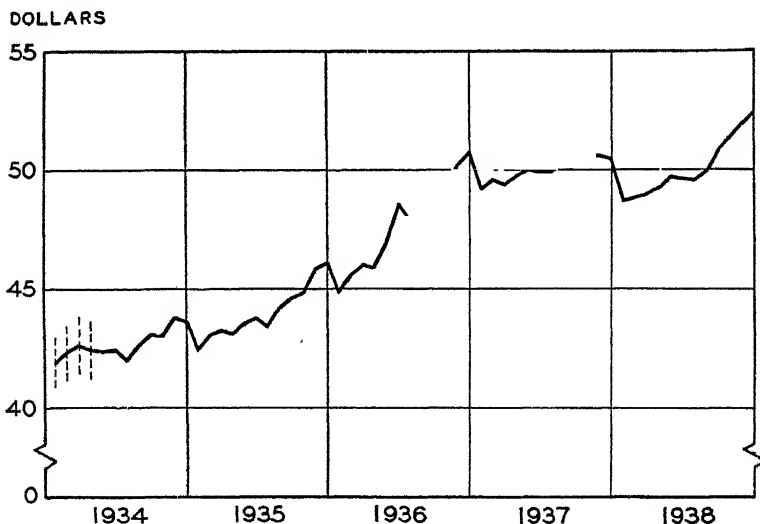
Chart 8.   Money in Circulation per Capita in the United States at the End of Each Month, 1934–1938.   (Data from Standard Statistics Company, *Basic Statistics*, p. A-22 and monthly supplements.)

Curves of frequency distributions. The curve of Chart 4 is a graphic representation of a frequency distribution. Frequency distributions will not usually continue into the second quadrant as does this one. In this instance, however, there were some negative incomes.

Table 14 shows a frequency distribution[2] of the grades of the 1937 graduating class of the United States Naval Academy. In order to show the

TABLE 14

FREQUENCY DISTRIBUTION OF
GRADES OF THE 1937 GRADUATING
CLASS OF THE UNITED STATES
NAVAL ACADEMY

Grade	Number of midshipmen
68.0–69.9	4
70.0–71.9	17
72.0–73 9	39
74.0–75.9	62
76.0–77.9	58
78 0–79.9	52
80.0–81.9	35
82.0–83.9	22
84.0–85.9	18
86.0–87.9	13
88.0–89.9	4
90.0–91.9	2
92.0–93.9	1
Total	327

Source. Table 25

genesis of the frequency distribution curve, we shall plot the data first as a series of rectangles or bars as in the "column diagram" of Chart 9. It will be noticed that the grades have been placed along the horizontal axis and the frequencies (number of midshipmen) along the vertical axis. There are as many columns in the chart as there were classes in the table, and the height of each column represents the frequencies in the corresponding class. This column diagram is transformed into a curve by connecting the midpoint of the top of each rectangle with the adjacent one, as shown by the dotted line in Chart 9. This is done upon the assumption that the frequencies in a class interval are evenly distributed throughout the class. The mid-value of a class is consequently taken as representing the class.[3] It will be observed that the dotted line cuts off

---

[2] Frequency distributions are discussed in Chapter VIII.
[3] This point is discussed at greater length in Chapter IX.
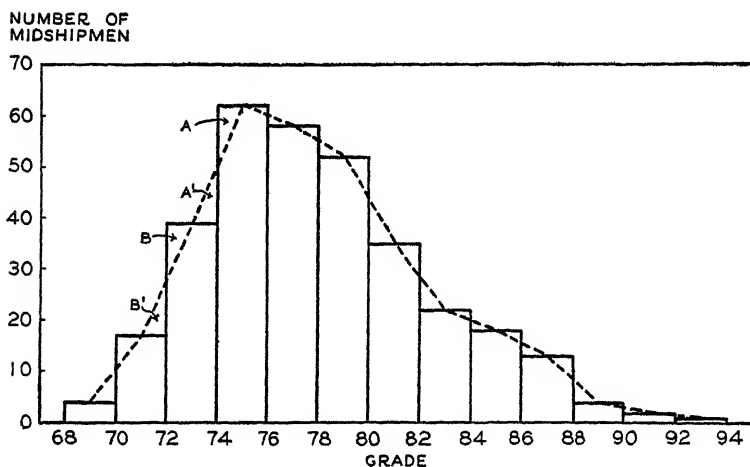
Chart 9.   Grades of the 1937 Graduating Class of the United States Naval Academy,
Shown by a Column Diagram and by a Frequency Curve.   (Data from Table 14.)

some small triangular pieces of the original rectangles and that it also includes some small triangles not formerly included, but it is obvious that triangle A = triangle A', triangle B = triangle B', etc.   Sometimes the curve is continued at each end to join the X-axis (indicating a frequency of zero) at the mid-value of the next possible class.   This procedure results in having the same area under the curve as is included in the rect-
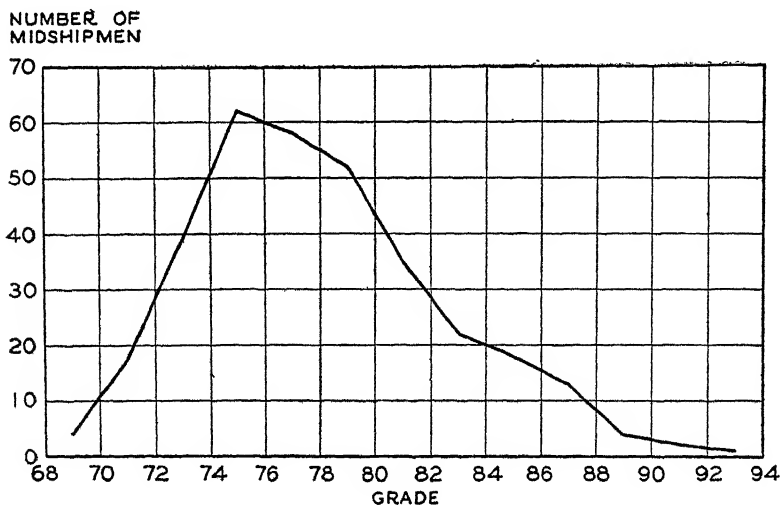
Chart 10.   Grades of the 1937 Graduating Class of the United States Naval Academy.
(Data of Table 14.)

angles. However, the result may sometimes be a curve which extends beyond zero on the X-axis and this is apt to be meaningless. The frequency distribution may be shown either as a column diagram or as a frequency curve (frequency polygon). The latter is more usual and the curve is plotted directly as in Chart 10, without the intermediate step of constructing columns.

In Table 15 the families in the United States have been classified according to the number of members in each. The data of this table could be

TABLE 15

Families in the United States,
by Size, 1930

Number of members	Per cent of all families
1	7.9
2	23 4
3	20 8
4	17.5
5	12 0
6	7.6
7	4.7
8	2.8
9	1.6
10	.9
11	.5
12 or more	.4
Total ..	100 0

Source. *Statistical Abstract of the United States, 1937,* p 50   T . ˆ: r(ˉ - ˉ ., ı ' famılıeˢ ʻ-ˢˏ ʾˏˉʳ-ˏ   ı '   . (ˏ '   ı ˉ clude   ' ˉ - .   ;.   ' - , - '.- lated   .   .   :- ˏ · or adoption, and groups sharıng tne same lıvıng accommodations as "partners "

plotted as a curve, similar in general plan to that of Chart 10. However, the classification into families of 1, 2, 3, 4, etc., members possesses a lack of continuity between the various classes, and the data are more properly shown by means of a bar chart similar in construction to that of Chart 124 (p. 291). The separation of the bars as in the latter chart emphasizes the lack of continuity between the categories.

Rules for Drawing Curves

While statisticians have not agreed upon a standard procedure setting forth in detail exactly how line diagrams should be constructed, there are certain rather obvious considerations of importance. The student who

is interested in going into more detail in regard to the technique of chart construction is referred to the books listed at the end of this chapter.
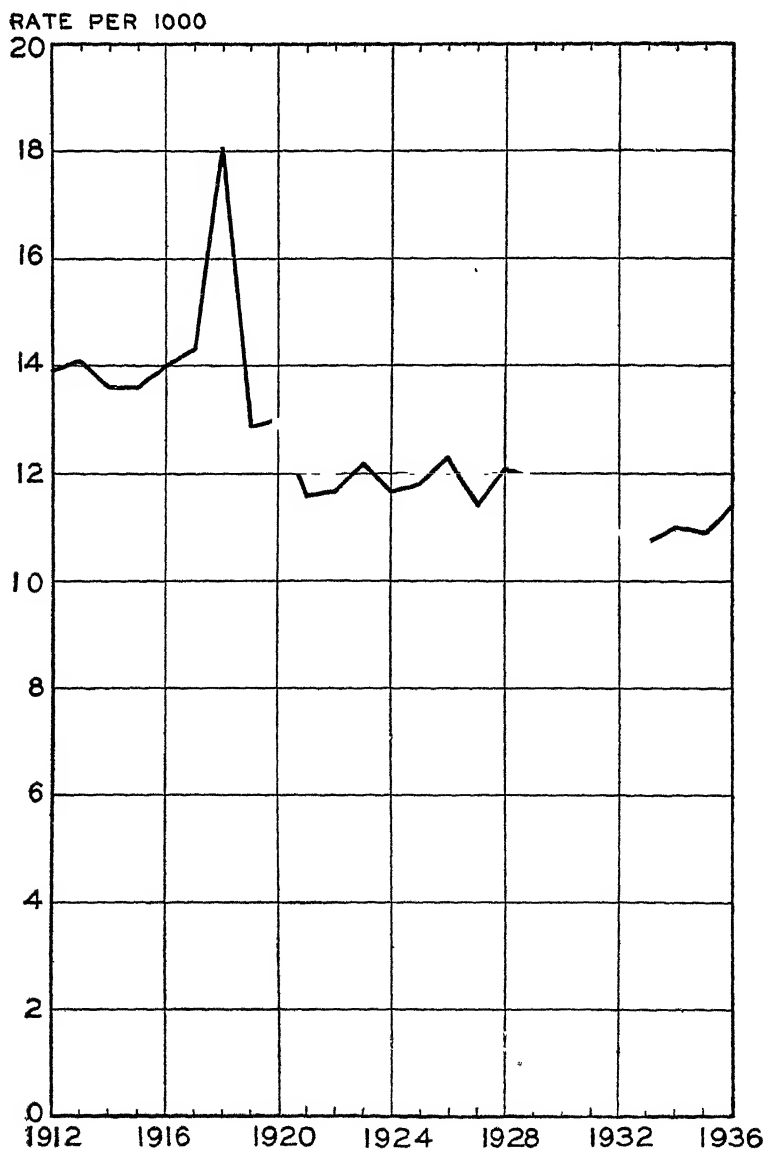
Chart 11. Death Rates per 1,000 Population in the Registration Area of the United States, 1912-1936. Data available just before going to press give the final rate for 1937 as 11.2 and the provisional rate for 1938 as 10.7. (Data from *Statistical Abstract of the United States, 1937*, p. 80, and Bureau of the Census, *Vital Statistics, Special Reports*, Vol. 4, No. 54.)

Zero on vertical scale.   The inclusion of a zero on the vertical scale of a curve is perhaps one of the most important rules.   Chart makers occasionally neglect to observe this principle and the result is always misleading, since the visual impression is incorrect.   In Chart 11, death rates in the United States from 1912 to 1936 have been plotted with reference to a vertical scale beginning with zero.   The same series of data appear in Chart 12, but on this chart the vertical scale begins at 10.   Chart 12 gives the reader a visual impression which is quite contrary to the facts.   The death
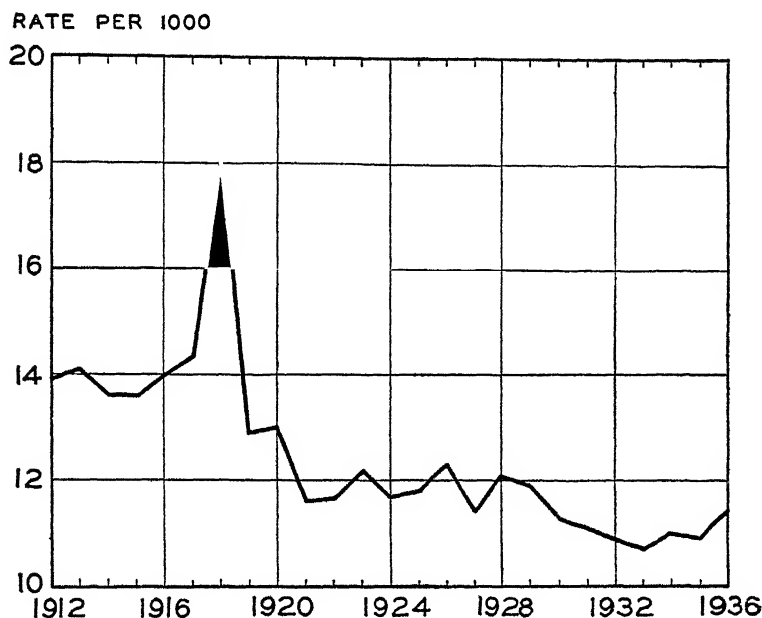
RATE PER 1000

Chart 12.   Death Rates per 1,000 Population in the Registration Area of the United States, 1912–1936.   This chart is incorrectly drawn, since the vertical scale begins with 10 and no clear indication of the omission of the zero is given.   (Data from same sources as Chart 11.)

rate in 1918 appears to have been about twice that for 1917, whereas Chart 11 shows that it was actually about 1⅓ times as large.   Again, in Chart 12 the death rate for 1928 *appears* to have been about 1½ times that for the previous year, whereas from the preceding chart it is clear that the 1928 rate was really not much greater than for 1927.   Very few readers notice the omission of zero on a vertical scale, and fewer still are apt to make due allowance for the omission in interpreting a curve.   It should not be necessary for a reader to refer to a scale in order to make approximate comparisons; the chart should be so drawn that comparisons may be

made visually and as quickly as possible.   Chart 12 also gives the impression that death rates have approached very closely to some sort of absolute lower limit.

There are several ways in which it is possible to show the zero (or to indicate clearly its omission), and also to avoid placing the curve high up on the chart.   Chart 13 shows a method in which a definite break is made across the chart.   Sometimes the parallel lines are serrated (notched) in-
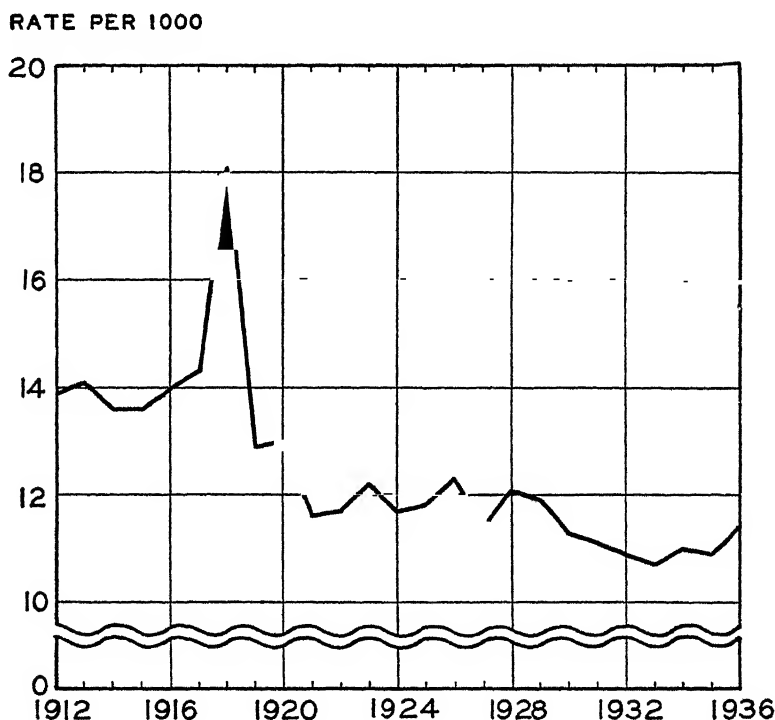
RATE PER 1000

Chart 13.   Death Rates per 1,000 Population in the Registration Area of the United States, 1912-1936.   (Data from same sources as Chart 11.)

stead of wavy.   They may be drawn freehand or, as in Chart 13, by making use of a bread knife as a ruler.   Charts 6, 8, 14, and 15 show other devices which are occasionally used.   Notice that Charts 6, 8, and 13 show the zero and a scale break, while Charts 14 and 15 do not show the zero but merely call attention to the fact that the vertical scale is incomplete.

Chart 16 appeared in the October 26, 1934, issue of *Railroad Data* and was also used as part of an exhibit in a hearing before the Interstate Commerce Commission.   The accompanying text indicated that the vertical

scale values were "pounds of coal required to move 1,000 gross ton miles of freight." Attention was directed to the increasing economy in the use of coal, but no warning was given as to the omission of the zero. The result is a misleading visual impression of a very marked increase in economy. It might, of course, be argued that this chart should not have zero for a base line, but rather a figure representing the minimum number of pounds of coal theoretically required to move 1,000 gross ton miles of freight under
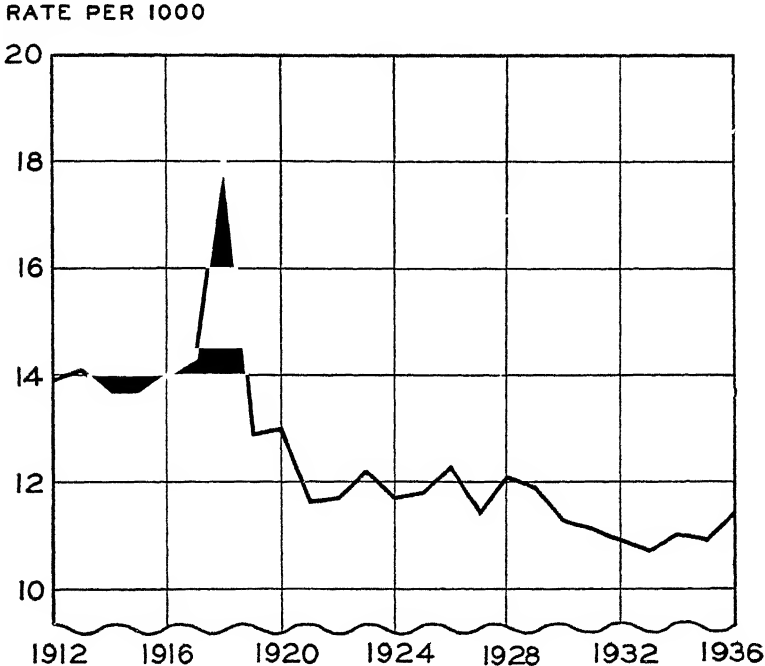
RATE PER 1000

Chart 14.   Death Rates per 1,000 Population in the Registration Area of the United States, 1912-1936.   (Data from same sources as Chart 11.)

ideal conditions. Nevertheless, the chart as it stands is subject to possible misinterpretation.

Occasionally curves will be seen lacking a zero on the vertical scale and purporting to show the growth of sales of a commodity or of membership in an organization. The omission of the zero makes the growth appear to be much more rapid than it really has been.

Chart 17 shows index numbers of the retail prices of food. This chart is unusual in two respects. In the first place it carries a zero for the vertical scale, which, though not wrong, is not necessary when price index numbers are being plotted, because it is hardly conceivable that prices

RATE PER 1000
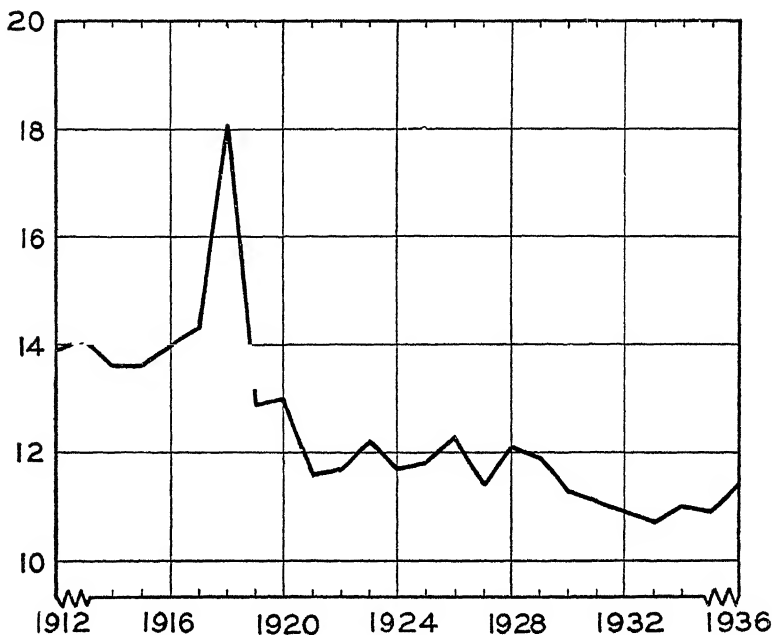
Chart 15. Death Rates per 1,000 Population in the Registration Area of the Unite
States, 1912-1936.   (Data from same sources as Chart 11.)
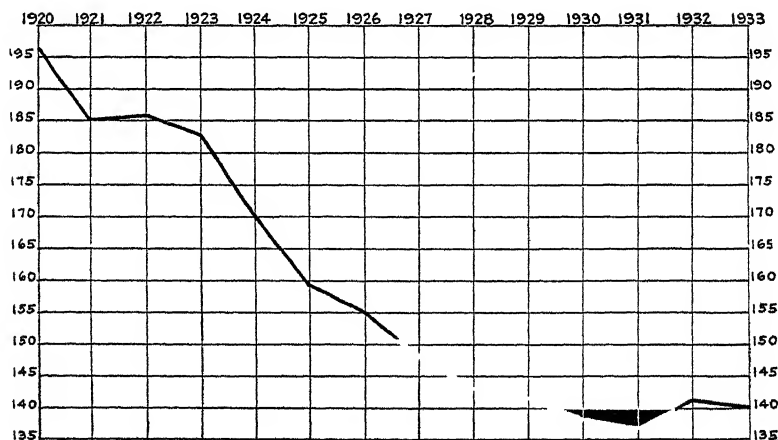
Chart 16.   Pounds of Coal Consumed per 1,000 Gross Ton-Miles of Freight Carried
by Class I Railroads in the United States, 1920-1933.   The vertical scale values are
pounds of coal required to move 1,000 gross ton-miles of freight, all fuels used having
been expressed in equivalent pounds of coal.   (From *Railroad Data*, October 26, 1934.)

will ever approach zero and because the base 100 is the basis of comparison.
The 100 line should always be emphasized when it is the base, as in this
chart.   Similarly the zero line should be emphasized, as in Chart 13, when
it is the base of the chart.   When charting index numbers, some persons
prefer to show the fluctuations above and below the base in terms of posi-
tive and negative values.   In the case of Chart 17, 100 would become
zero, 125 would become +25, and 75 would become −25.   The vertical
scale of Chart 17 would be altered to read +75, +50, +25, 0, −25, −50,
−75, −100.   The curve itself would remain unchanged.   The second
unusual feature of Chart 17 is the treatment of the horizontal and vertical
guide lines, which results in giving the curve an unusually clear profile.
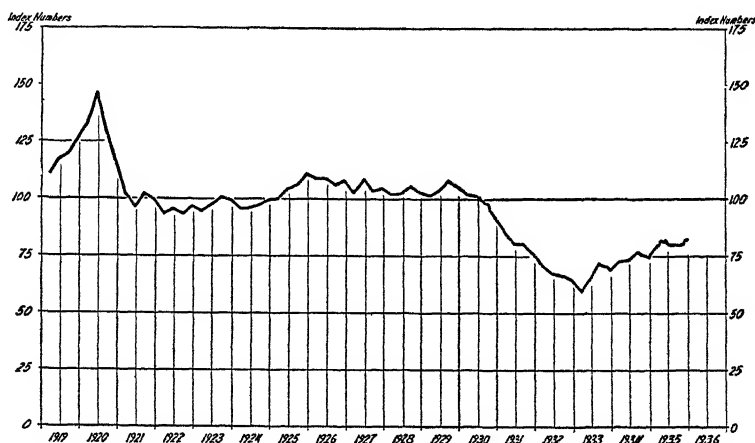
Chart 17.   Retail Cost of Food in the United States, 1919-1935, 1923-1925 = 100.
(From *Monthly Labor Review*, February 1936, p. 495.)

Notice also that space has been left to add later data.   This practice allows
the same original chart to be reproduced month after month by merely
extending the curve as new data become available.

Ruling curves.   The curve or curves representing the data should stand
out clearly from the background of the chart.   The curve should therefore
be ruled more heavily than the coordinates.   (When two or more curves
are shown which follow each other closely, it is sometimes necessary to
use lightly ruled lines for the curves.   See, for example, Chart 187, p. 507.)
As will be seen from the various charts, the plotted points are not usually
shown since the attempt is to show the general situation rather than the
individual readings.

When several curves are drawn on the same axis, it is essential that
each curve stand out clearly.   Thus we may use solid, dotted, and dashed

lines, and we may use heavy and light lines. If a light line is used for a curve it should ordinarily not be so light as the coordinates. The suggested rulings are listed below as A and B.

A	B	C
————————	————————	—o—o—o—
••••••••••••••••	--------------------	—•—•—•—
— — — — — —	— — — — — —	⊙⊙⊙⊙⊙
A. These lines are recommended if not more than three curves are to be drawn.	B. If more than three curves are to be drawn, these lighter lines may be used.	C. These lines are not recommended unless plotted points are to be indicated by means of the circles or dots.
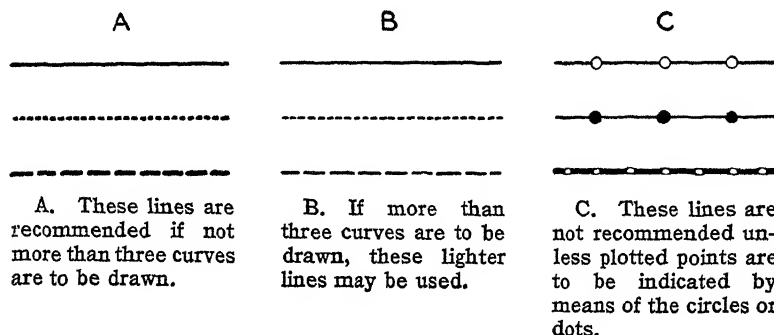
When two or more curves appear on a chart, each should be clearly identified. This may be accomplished, preferably, by labeling the curves as in Chart 28, or by means of a legend as in Chart 23.

It is ordinarily well to avoid the use of more than two or three curves on one chart. Particularly if they cross and re-cross, confusion is likely to result. When several curves appear on a large wall chart which is to be presented to a group, different colors may occasionally be used, though it is usually better practice to reserve the use of color for those occasions when special emphasis is to be placed on one or two curves. Black, red, green, light or medium blue, and medium or dark orange are readily distinguished If there is a likelihood that the wall chart is to be photostated, photographed, or reproduced for printing, black and red may be used in solid and broken, light and heavy, combinations since the red line will reproduce as black. Blue, yellow, and some shades of green photograph either not at all or faintly.

Coordinates. Chart makers emphasize the zero line by making it a little heavier than the other marginal lines. In similar fashion a 100 per cent line (or other base with which comparisons are made) may be stressed. The marginal vertical and horizontal lines may be made slightly heavier than the other coordinate lines.

The coordinate lines should be drawn very lightly. No more coordinate lines should appear than are necessary to assist in reading the chart. Occasionally all coordinates are omitted, as in Chart 5. If it is desired to have a closely ruled grid in order to make plotting easy, an effaceable ruling may be had. The coordinates of this paper may be washed off after

they have served their purpose.[4]   When a chart is to be reproduced, a closely ruled grid of light blue may be used.   The lines which should appear in the reproduction are ruled in black.   The blue lines of the background do not show up in the reproduction under ordinary conditions. Most of the charts in this text were drawn on such a light blue background.

In order to insure a proper understanding of a chart, the two scales should be clearly labeled.   Not only should the nature of the variable be indicated, but the unit used should also be stated.   Note, for example, in Chart 4 the horizontal axis shows incomes, the unit being thousands of dollars.   Chart 16, however, has no label on the vertical scale; it is necessary to read the accompanying text in order to understand the chart. Occasionally a curve of a long time series may be rather extended horizontally.   In such instances it is sometimes desirable to repeat the vertical scale at the right of the chart, as in Chart 17.

Chart proportions.   It is hardly possible to give an objective rule as to the proper proportions for a curve diagram.   It should be noted, however, that bizarre impressions result from over-expanding or over-contracting either scale used for a curve. In Chart 18 the vertical scale is exaggerated in relation to the horizontal scale; in Chart 19 the horizontal scale is exaggerated.   The former gives an impression of tremendous fluctuations; the latter conveys the idea that
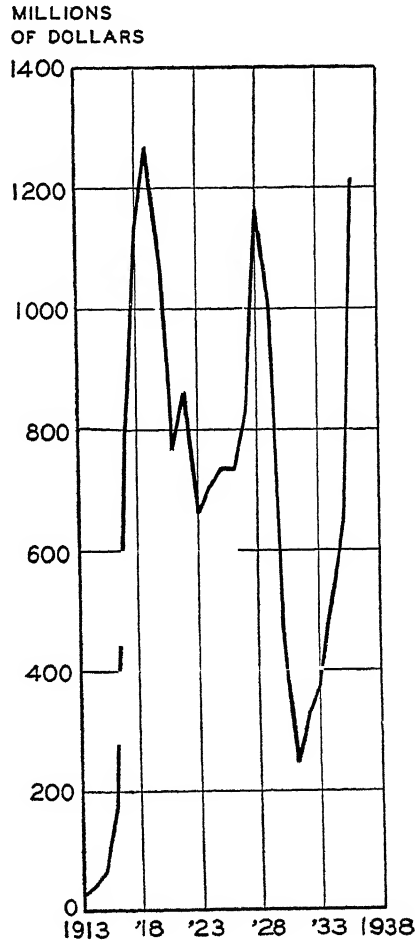
Chart 18.   Individual Federal Income Taxes in the United States, 1913-1936.   The vertical scale is exaggerated.   The use of '18, '23, etc. is ordinarily to be avoided.   (Data from same source as Chart 3.)

---

[4] From Carl Schleicher and Schull, 167 East 33rd Street, New York City.   Since no ink is entirely waterproof, best results will be obtained by drawing the chart in hard pencil, washing off the coordinates, and then inking in the lines for the finished chart.

MILLIONS OF DOLLARS
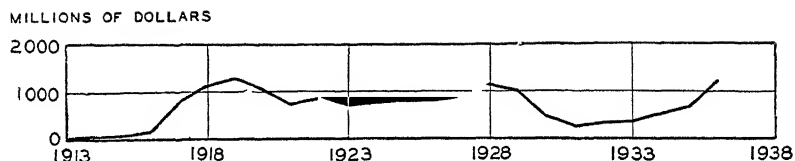
Chart 19. Individual Federal Income Taxes in the United States, 1913-1936. The horizontal scale is exaggerated. (Data from same sources as Chart 3.)
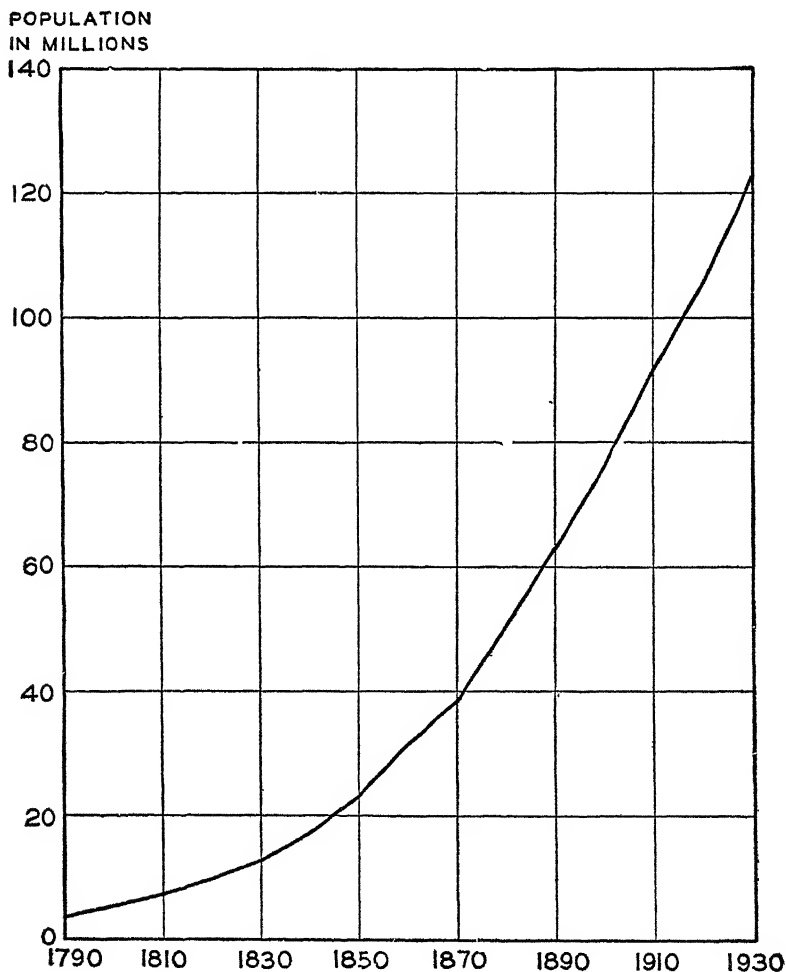
Chart 20. Population of the United States, 1790-1930. (Data from *Fifteenth Census of the United States, 1930*, Population Volume I, pp. 10–11.)

the level of individual income tax payments has undergone relatively un-important fluctuations. These two charts indicate distorted results of replotting the data shown properly in Chart 3. Rules of thumb are often un-satisfactory because they are apt to be adopted blindly. However, it has been suggested that the proper proportions are those which result in a 45 degree angle for the movements of the curve which are to be emphasized.

Chart 20 shows the population of the United States from 1790–1930. Charts 21 and 22 show the same data; however, Chart 21 gives a visual impression of rapid growth, while Chart 22 gives a visual impression of slow growth. A chart could, of course, be made to show amount of growth (or per cent of growth) for each census in relation to the preceding. In this case the vertical scale would show "amount of increase over preceding census" (or "per cent of increase over preceding census").

Although the two charts just referred to are curves of time series, it should be under-stood that the same false impressions are given by frequency curves if one scale is over-expanded.

Lettering. All lettering on a chart, in-cluding scale labels, scale values, legend, curve labels, and any other words or figures should be placed horizontally, if possible. Occasionally space limitations may necessi-tate placing the vertical scale label in a vertical position, as in Chart 23, but such a limitation is not often present. Needless to say, all lettering should be legible. Free-hand words and figures may be made very attractive when executed by a skilled per-son (see, for example, Chart 135). The amateur, may however, make excellent formal letters and figures (with a little prac-tice) by the use of stencil lettering devices

POPULATION IN MILLIONS
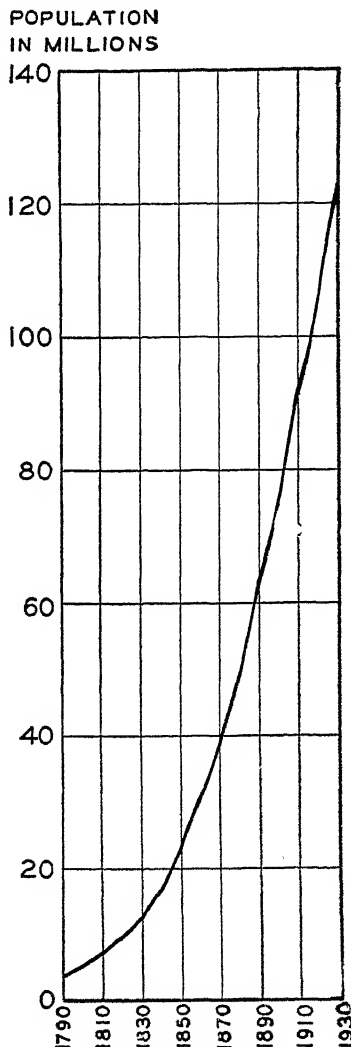
Chart 21. Population of the United States, 1790–1930. The horizontal scale is poorly chosen in respect to the vertical scale. It is usually better to avoid writing the horizontal scale values in the man-ner shown here. (Data from same source as Chart 20.)

available from artists' or draftsmen's supply houses. Nearly all of the charts in this text, except those reproduced from other publications, were lettered by means of such devices. The lettering of Charts 39 and 62 and the inserts on a number of other charts (for example, Charts 4, 88, 89, 90, 91, and several in Chapter XII) was done by means of a Vari-Typer composing machine, which is essentially a typewriter using many different kinds of type.[5]

POPULATION
IN MILLIONS
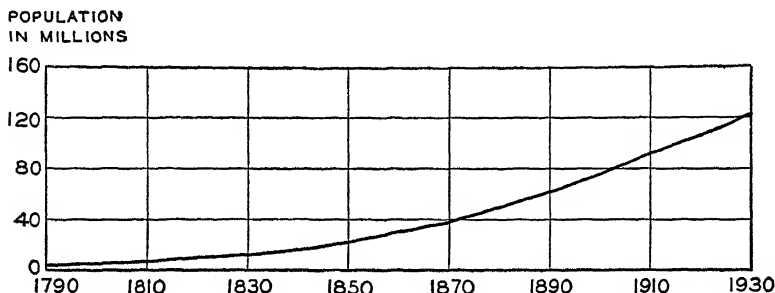
Chart 22. Population of the United States, 1790-1930. The vertical scale is poorly chosen in respect to the horizontal scale. (Data from same source as Chart 20.)
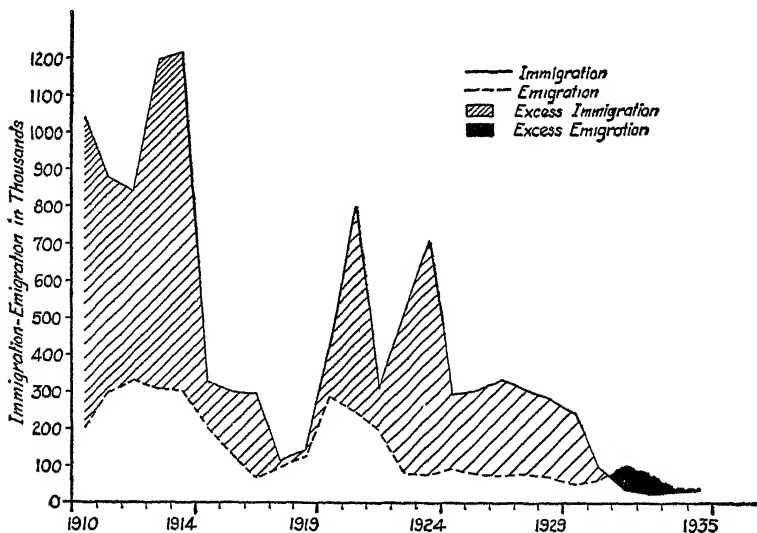
Chart 23. Immigration to and Emigration from the United States, 1910-1935. (From "Immigration Becomes Emigration," by Robert E. Chaddock in *The Independent Journal of Columbia University*, March 6, 1936.)

_____

[5] The stencils are available from Keuffel and Esser Co., 127 Fulton Street, New York City, and from Wood-Regan Instrument Co., Nutley, New Jersey. The Vari-Typer is sold by the Ralph C. Coxhead Corporation, 17 Park Place, New York City.

Title. Each chart, like each table, should have a title, which should state clearly and succinctly what the chart purports to show. The title of a printed chart may appear either above or below the chart, but is usually below. The titles of large wall charts are often placed above the grid.

Source. Again, as in the case of a table, each chart should contain a source reference to indicate the author, title, volume, page, publisher, and date of the publication from which the data were taken. Naturally the cautions regarding comparability of data taken from the same source or different sources, mentioned in Chapter II, apply with full force to the figures used for making charts.

Line Diagrams for Special Purposes

Net balance charts. Chart 5 shows one method of indicating the net total of two series. For each of the years, total emigration was subtracted from total immigration and the result plotted as a positive or negative
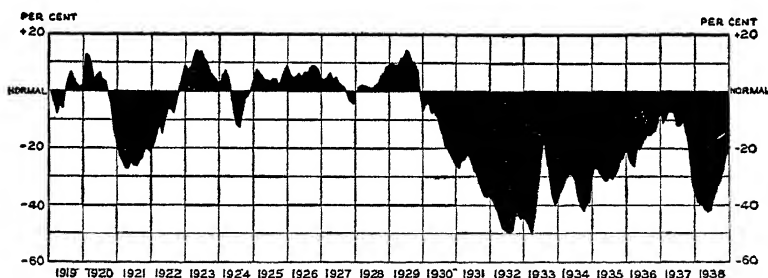
Chart 24. Cleveland Trust Company Index of Business Activity in the United States, 1919-1938. The complete series runs from 1790 to date. (Data from various issues of The Cleveland Trust Company, *Business Bulletin.*)

figure. The balance of trade (value of exports minus value of imports) may be shown in the same manner, as may also profit and loss. An alternate method of showing the migration data is illustrated in Chart 23. Here the curves of immigration and of emigration are given; and the excess of immigration is indicated by the height of the shaded area, while the excess of emigration is shown by the height of the black portion.

Silhouette charts. Chart 23 (referred to in the preceding paragraph) illustrates not only the showing of net amounts rather than gross amounts, but likewise the practice of shading the area between the curves in order to obtain emphasis. Chart 24 is similar to Chart 5 in that it shows fluctuations above and below a base line. In Chart 24, however, the areas of the curve have been emphasized by filling in with black. The result is a more

striking portrayal of the "plus" and "minus" parts of the curve.   A chart of this type is even more effective when the "plus" areas are filled in with black and the "minus" areas are filled in with red.   Another example of a silhouette chart is shown by Chart 255, page 815.

Maximum variation charts.   The Library of Columbia University displayed in an illuminated glass case a number of valuable old prints.   For the proper preservation of the prints it was desired to maintain the temperature between 70 and 80 degrees Fahrenheit.   The problem consisted of adjusting radiation of heat from the case, ventilation and conduction, and the proximity to nearby radiators so that the temperature inside the case would remain within the desired limits.   A recording thermometer was placed in the case and the temperature was continuously recorded over
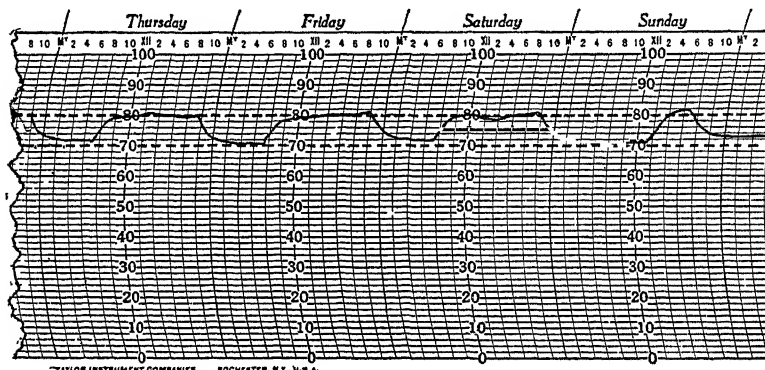
Chart 25.   Temperature Fluctuations in a Library Display Case.   Temperature is in degrees Fahrenheit.   The curved ordinates are made to correspond to the arc described by the recording pen of the thermometer.   (From the library of Columbia University.)

an extended period.   In Chart 25 a four-day section of one of the charts is shown.   During these days there was no heat in the adjoining radiator, and it may be seen that the temperature never fell below 70 degrees but did slightly exceed 80 degrees on several occasions.   On Thursday, Friday, and Saturday the library was open to the public from 8 a.m. to 10 p.m.; on Sunday, from 2 to 6 p.m.   The dashed lines have been added by the authors and serve to stress the limits beyond which the temperature should not fluctuate.

Range charts.   Chart 26 shows a device by means of which the range of stock prices may be depicted.   It will be noticed that the black band expands when the range is greater and contracts when the range is smaller. The white line indicates the closing price.   An alternate method of showing the same figures is illustrated in Chart 27.   Here the top of each bar represents the high for the day, while the bottom of each bar represents
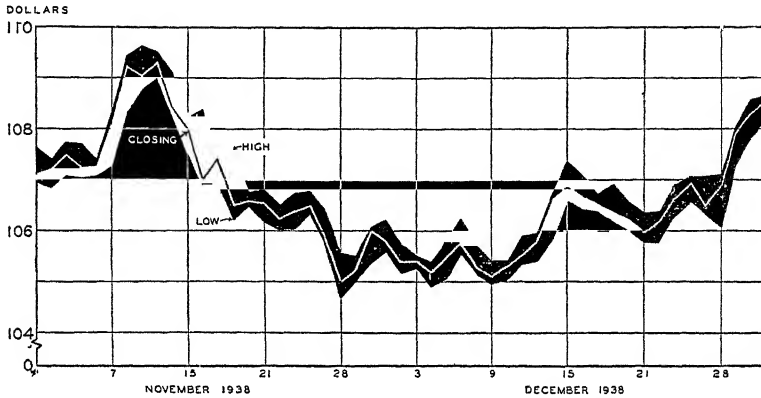
Chart 26.   High, Low, and Closing Prices of 100 Stocks as Shown by the New York *Herald Tribune* Averages, November and December, 1938.   Data are plotted for only those days during which the stock exchange was open.   (Data from New York *Herald Tribune*)

the low for the day.   The line connecting the bars represents the closing price.   Charts such as these may be used for showing commodity prices and other sorts of data if it is desired to show a range of variation over a period of time.

Z charts.   The *Z* chart consists of three curves on the same axes as shown in Chart 28.   Usually the chart covers a period of one year, by months.
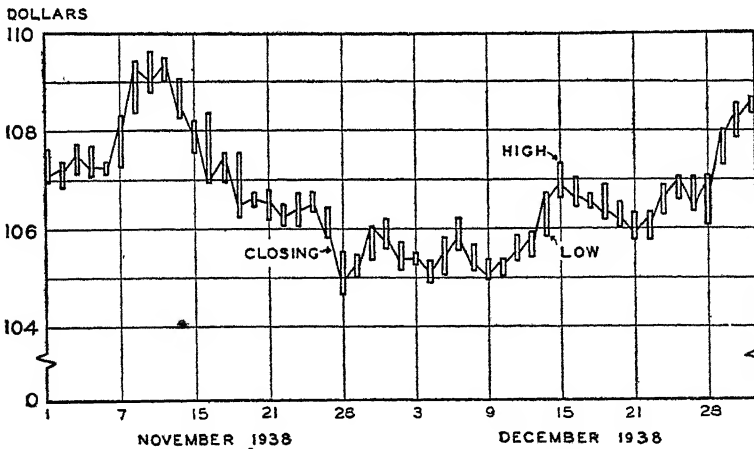
Chart 27.   High, Low, and Closing Prices of 100 Stocks as Shown by the New York *Herald Tribune* Averages, November and December, 1938.   Data are plotted for only those days during which the stock exchange was open.   (Data from New York *Herald Tribune*.)

One curve shows the monthly figures, another shows the cumulative figures from the beginning of the year, while the third shows the total for the twelve months ending with each month. This last curve is generally called the *moving annual total* curve; more specifically it is a 12-month moving total for the twelve months ending with each specified month. Two vertical scales are used with the Z chart since, if the monthly data were plotted against the same scale as the other data, the fluctuations of the monthly data would not be apparent. The Z chart is often used for internal business purposes, showing, for example, data of production and sales. There is no reason why it should not be more widely used for showing other types of data such as those shown in Chart 28. It is, of course,
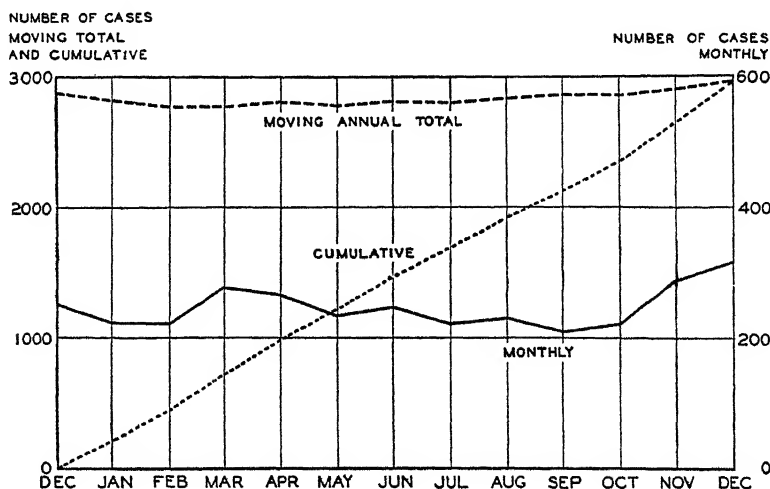
Chart 28. Intake of Cases New to a Family Service Agency: Monthly, Cumulative, and Moving Annual Total, 1937. (Data from Welfare Council of New York City.)

limited to those situations in which the chart maker is interested in visualizing: (1) the figure for a given month, (2) the figure for each month for that part of the calendar (or fiscal) year which has elapsed, and (3) the figure for the twelve months ending with each given month.

Except for special purposes such as this, it is not usually desirable to use two, or more, vertical scales (sometimes referred to as "multiple scales") on a chart of the type described in this chapter. The occurrences of fluctuations (but not their magnitudes) in two series expressed in different units may occasionally be compared on a chart having two different vertical scales. However, the use of two, or more, different vertical scales is likely to give false visual impressions of the comparative magnitudes of changes occurring in the various series.

Varying horizontal scale charts. Occasionally it is desired to show annual data over a number of years, and monthly data for one or two more recent years. This may be done as in Chart 29, in which the horizontal scale is expanded to show the monthly data in more detail. Notice that the two parts of the chart are separated by a break. Similarly, a change in horizontal scale may be in order if we wish to show a combination of annual or monthly data with weekly data, or a combination of annual, monthly, or weekly data with daily data.

Multiple axis charts. Occasionally it is desirable to compare the fluctuations of several curves and yet to have each curve stand out clearly. A simple method of accomplishing this result is to plot the different
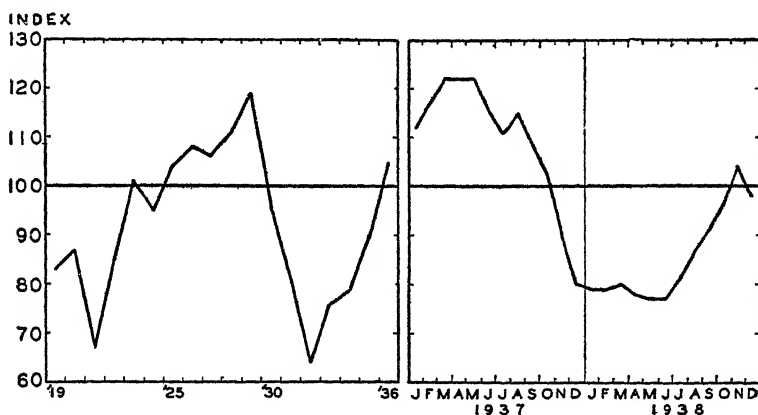
Chart 29. Federal Reserve Board Index of Industrial Production in the United States. Annually 1919–1936 and Monthly 1937 and 1938. (Data from *Federal Reserve Bulletin*, January 1939, p. 62, and press releases; 1923-1925 = 100.)

curves along different horizontal axes, these different X-axes being arbitrarily separated by convenient vertical distances. An illustration is Chart 176. Here the different curves have been brought close together for ease of comparison, but there is no crossing of the lines. Although different horizontal axes are employed, the vertical and horizontal scales remain the same. In interpreting such a chart on arithmetic graph paper (as distinguished from semi-logarithmic graph paper described in the following chapter), it should be remembered that the comparison afforded is that of absolute and not of relative changes. It is unlikely that the use of this type of chart will be found desirable for presentation to the general reader, unless the diagram is accompanied by a clear explanation.

Component part charts. Chart 30 shows the number of persons in the United States at each census from 1850 to 1930, in each of four general age

MILLIONS
OF PERSONS
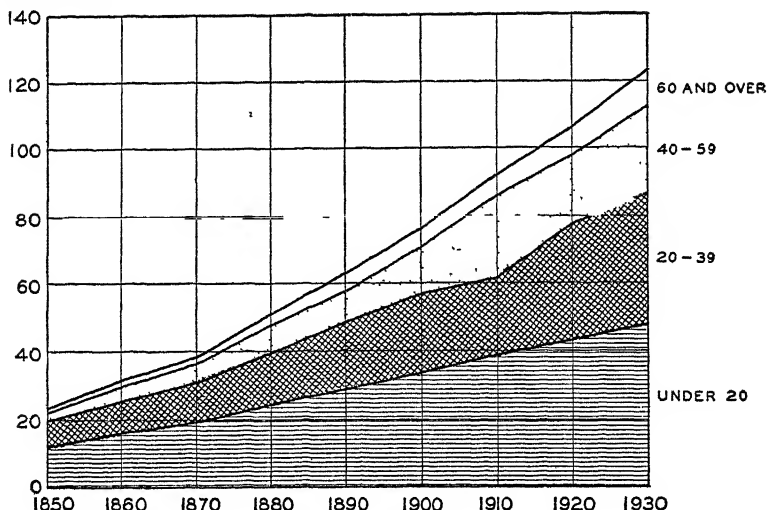
Chart 30. Population of the United States in Each Specified Age Group, 1850-1930. (Data from *Fifteenth Census of the United States, 1930*, Population Volume II, p. 576.)

PER CENT
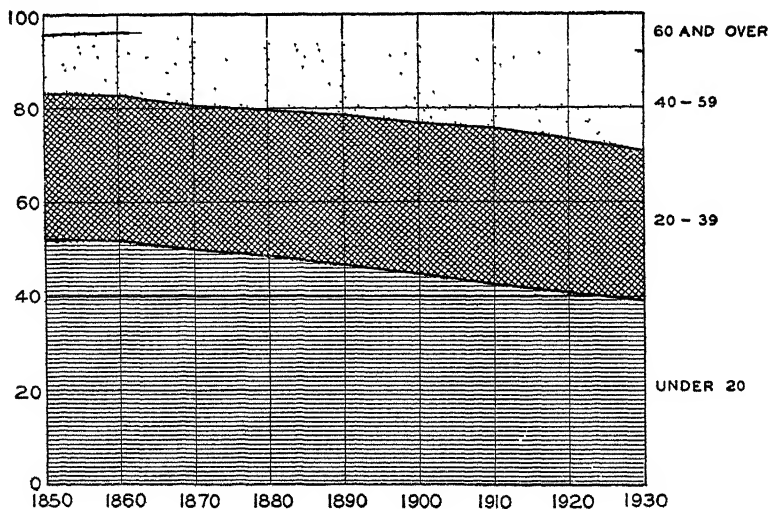
Chart 31. Proportion of the Population of the United States in Each Specified Age Group, 1850-1930. (Data from *Fifteenth Census of the United States, 1930*, Population Volume II, p. 576)

groups.   The width of each band indicates the number of each age in the country at a given census.   It is possible to observe, from this type of chart, whether or not a given group is increasing or decreasing, and whether or not the total of all groups is increasing or decreasing.   The *relative* importance of a particular group cannot be visualized from Chart 30, but in Chart 31 the age groups are shown according to the proportions which they constitute of the total population   Here it may be clearly seen that there has been a decrease in the proportion of younger persons and an increase in the proportion of older persons in the population.   When component part data covering a few years are to be shown graphically, a bar
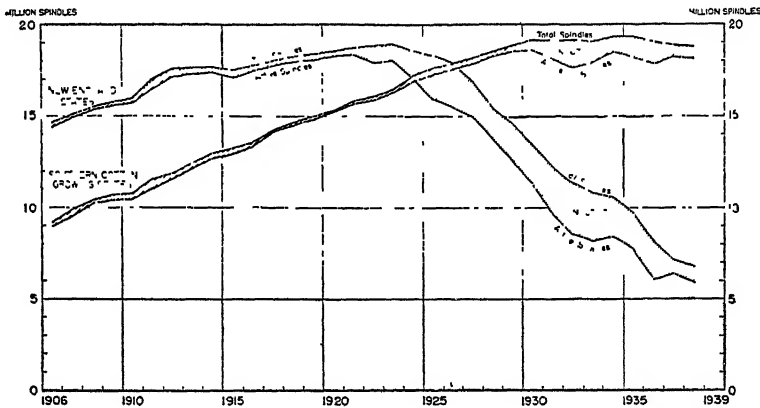
Chart 32. Total, Active, and Inactive Cotton Spindles in the New England and Southern Cotton Growing States, 1906-1938.   (From *Monthly Labor Review*, December 1938, p. 1245.)

chart such as Chart 71 or 72 may be used.   When a number of years are to be shown, the general trend can be more easily pictured by curves.

Chart 32 is another illustration of a component part chart.   In this chart one part, inactive cotton spindles, is emphasized by shading.   The data of total spindles and of active spindles are plotted above the zero line. The difference between these two is the shaded area, inactive spindles.

Frequency distribution and range chart.   In 1928 the United States Personnel Classification Board made a study of the white-collar workers employed by the Federal Government outside of Washington and exclusive of postal employees.   One phase of the investigation involved the collection of data concerning the salaries paid in comparable occupations in private employment.   Chart 33 shows a method by which comparison was made between the distribution of salaries paid to one group of employees in

private concerns (referred to as "general industry" on the chart) and the salary range for similarly qualified persons employed by the government. The source from which this chart was taken shows a number of such diagrams, for various occupations and occupation groups.   In some instances the government was paying more than the usual rate in private employment; in some instances the government rate was below the going private rate; in some instances there was substantial agreement; in every chart shown, however, the *range* of salaries paid by the government was narrower than in private employment.   This may reflect a more exact classification of employees than in the aggregate of private concerns studied
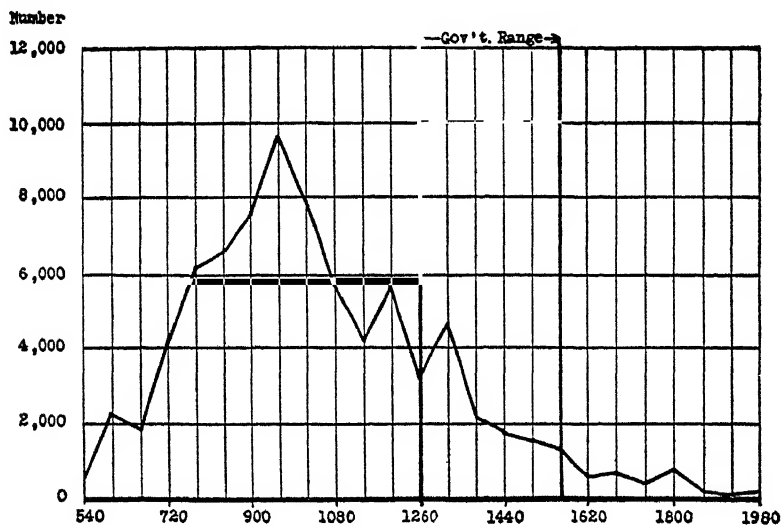
Chart 33.   Annual Salaries Paid to Employees in General Industry Corresponding to the CAF-1 Classification and Government Field Service Range of CAF-1 Salaries, 1928. The CAF-1 grade is the lowest grade of worker in the "clerical, administrative, and fiscal service."   It includes adding machine operators, duplicating machine operators, addressing machine operators, file clerks, punch card operators, routine typists, and so forth.   (Reproduced from United States Personnel Classification Board, *Report of Wage and Personnel Study*, House Document No. 602, 70th Congress, 2nd Session, p 198.)

Instead of comparing a frequency curve of private salaries with a range for government salaries, as in Chart 33, a comparison of two frequency curves could have been shown. This sort of diagram is discussed in Chapter VIII.   It possesses the advantage of showing not only the range of the second set of data, but also its distribution within that range.

Selected References

H. Arkin and R. R. Colton· *Graphs· How to Make and Use Them*, Chapters I–IV. Harper and Brothers, New York, 1936.   Chapter III discusses equipment;

W. C. Brinton: *Graphic Methods for Presenting Facts*, Chapters V–IX; McGraw-Hill Book Co., New York, 1914.   Examples from many sources.

W. C. Brinton. *Graphic Presentation*, Chapters 33–37; Brinton Associates, 608 West 45th St., New York, 1939.

A. C. Haskell: *Graphic Charts in Business* (Second Edition), Chapters I–IV, VIII, XII; Codex Book Co., Norwood, Mass., 1928.

K. G. Karsten: *Charts and Graphs*, Chapters XV–XX, XXII; Prentice-Hall, Inc., New York, 1923.

B. D. Mudgett. *Statistical Tables and Graphs*, pages 90–145; Houghton Mifflin Co., Boston, 1930.

J. R. Riggleman. *Graphic Methods for Presenting Business Statistics* (Second Edition), Chapters I, II, IV, VI, Appendix; McGraw-Hill Book Co., New York, 1936.   Chapters I and II provide an introductory treatment of charts.   The appendix deals with drawing and lettering.

Subcommittee on Preferred Practice in Graphic Presentation· *Code of Preferred Practice for Graphic Presentation: Time Series Charts*, American Society of Mechanical Engineers, New York, 1936.   From an engineering point of view.

CHAPTER V

GRAPHIC PRESENTATION

THE SEMI-LOGARITHMIC OR RATIO CHART

Absolute and Relative Growth

When considering the development of a series of statistical data over a period of time, we are sometimes interested in the *amount* of growth that has taken place, but more often we wish to know something about the *relative* growth or *rate* of change.[1]   Diagrams such as Charts 3, 8, and
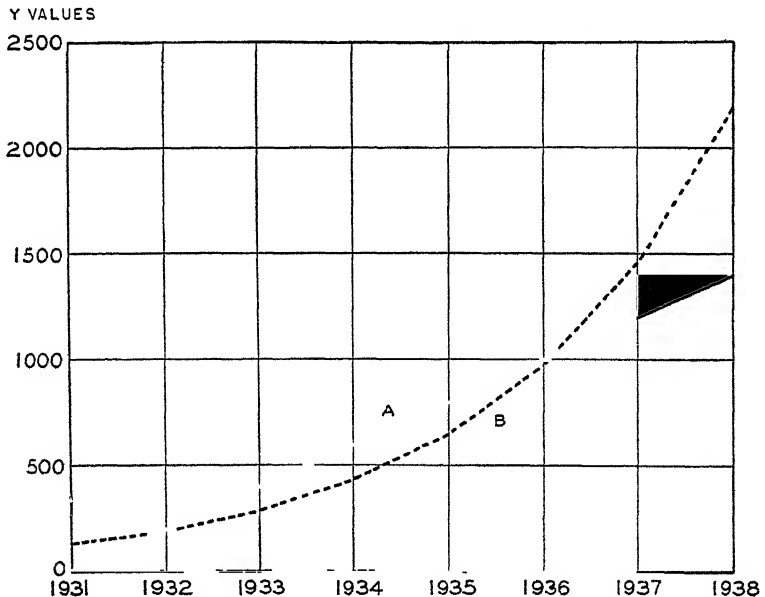
Chart 34.   An Arithmetic Progression (A) and a Geometric Progression (B) Plotted on an Arithmetic Grid.   (Data of Tables 16 and 17.)

---

[1] The terms "relative" and "rate" are susceptible of various meanings.   In this book, "relative change" (increase or decrease) and "rate of change" (increase or decrease) are used to mean change in relation to the value of the same series at a preceding date. This is sometimes referred to as "rate per cent of change" or "percentage rate of change." The change would be "per annum" if the comparisons were yearly.

various others in Chapter IV are of the familiar type, having what are termed arithmetic scales, and are of use, primarily, for indicating absolute changes in the factor shown on the Y-axis. It is the purpose of this discussion to explain a slightly different sort of grid which enables one to visualize the relative change that is taking place in a plotted series.

The ability of the usual type of chart to give a satisfactory visual impression of absolute, but not of relative, changes is brought out by Chart 34. Curve A represents a constant amount of increase of 200 units per year (see Table 16), and this, or any other, *arithmetic progression* (constant amount of increase or decrease) will be depicted by a straight line when plotted on the conventional or arithmetic grid. Curve B, however, is the result of plotting a series of figures which begin with 128 and increase 50 per cent each year (see Table 17). It will be noticed that this curve is not a straight line; the curve bends upward more and more sharply as time passes.

TABLE 16

AN ARITHMETIC PROGRESSION

Year (X value)	Y value	Amount of increase
1931	0	.
1932	200	200
1933	400	200
1934	600	200
1935	800	200
1936	1,000	200
1937	1,200	200
1938	1,400	200

TABLE 17

A GEOMETRIC PROGRESSION

Year (X value)	Y value	Per cent of increase
1931	128	. . .
1932	192	50
1933	288	50
1934	432	50
1935	648	50
1936	972	50
1937	1,458	50
1938	2,187	50

A series showing a constant rate of increase or decrease is known as a *geometric progression*, and any geometric progression will yield a curved

line when plotted on an arithmetic grid.[2]   A geometric increase is represented by a curve which slopes upward and is concave upward, as in Curve *B* of Chart 34; a geometric decrease is represented by a curve which slopes downward and is concave upward.   A serious difficulty in interpreting such curves, however, lies in the fact that the eye cannot discern whether or not a particular curved line does or does not represent a constant rate of change.   Chart 35 depicts a series which is neither an arithmetic nor a geometric progression.   The data of Table 18 show that the series increases

Y VALUES
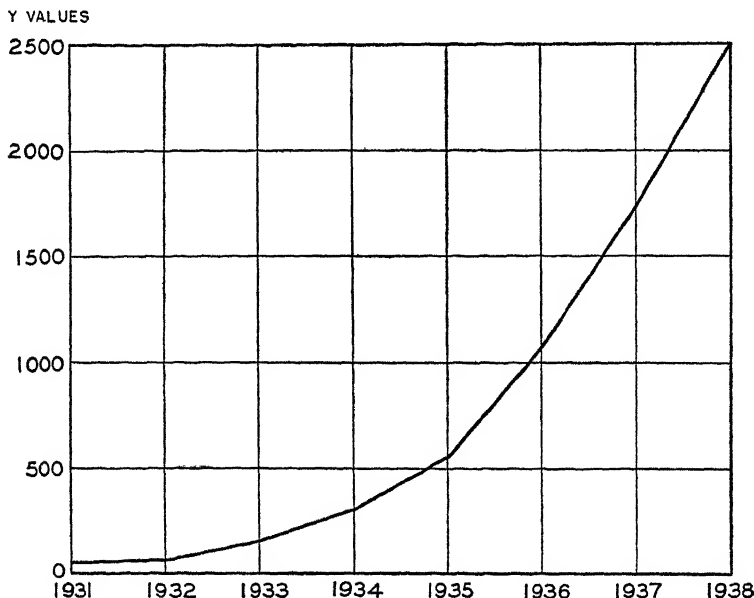
Chart 35.   A Series Increasing by Increasing Amounts.   This series is not a geometric progression, but may give that visual impression.   (Data of Table 18.)

more rapidly than an arithmetic progression, and the eye can grasp this fact because the curve bends upward.   The table also indicates that the series increases at a rate which is not constant.   Visually, however, this fact is not apparent.   It is not possible for the reader of an arithmetic chart to be sure whether a curved line such as this represents a constant rate of increase, a rate of increase which is diminishing, or a rate of increase which is accelerating.   Any series of figures that increases more rapidly

[2] A curve representing a geometric progression is termed an "exponential curve" and is indicated by the equation $Y = ab^X$.   The reader may be familiar with this equation in the form $P_n = P_0(1 + r)^n$, which is the compound interest equation and is discussed on pp. 225–226.   A straight line representing an arithmetic progression is indicated by $Y = a + bX$.

than an arithmetic progression (for example, 10, 12, 15, 19, 24, 30) slopes upward and is concave upward when plotted on an arithmetic grid; any series of figures that decreases less rapidly than an arithmetic progression (for example, 100, 91, 83, 76, 70, 65) slopes downward and is concave upward when shown on arithmetic coordinates.

TABLE 18

A SERIES OF INCREASING VALUES

Year (X value)	Y value	Amount of increase	Per cent of increase
1931	50	.	
1932	80	30	60.0
1933	160	80	100.0
1934	300	140	87.5
1935	550	250	83.3
1936	1,080	530	96.4
1937	1,730	650	60.2
1938	2,500	770	44.5

Before proceeding to develop the basis for the semi-logarithmic or ratio grid, which will enable us to visualize rates of change, let us examine further the arithmetic grid. Chart 36 shows the growth of motor vehicle registrations in the United States and in Canada from 1917 to 1938. We can see from this chart that registrations in the United States increased rapidly and, apparently, in approximately an arithmetic progression from 1917 to 1929; held fairly constant from 1929 to 1930; dropped in 1931, 1932, and 1933; and resumed the upward movement from 1934 to 1937, only to fall slightly in 1938. Changes in registration in Canada are difficult to see because the scale which must be used to accommodate the United States causes the curve for Canada to fall rather close to the base line. However, it appears that registrations in Canada increased from 1917 to 1930; decreased in 1931, 1932, and 1933; and increased in the five following years. It is quite obvious that from 1917 to 1929 the *amount* of increase each year was greater for the United States than for Canada, but there is no way of knowing from the appearance of the curves which country had the greater *relative* increase.

It would not do to replot the data of Chart 36 by using one vertical scale for the United States and another for Canada, in order to magnify the movements of the curve for the latter. The fact that one curve is below another on an arithmetic grid tells us at a glance that the lower curve represents a series of smaller magnitude than does the upper. If two vertical scales are used, we have really two distinct, non-comparable charts, and

no *satisfactory* visual comparisons may be made in respect to (1) the size of the two series plotted, (2) the amount of change which has taken place in one series in comparison with the amount of change in the other, or (3) the rates of change of the two series.

A Grid to Show Rates of Change

From what has already been said it must be obvious that graphic comparisons in respect to rates of change will be facilitated if we can employ

MILLIONS OF VEHICLES
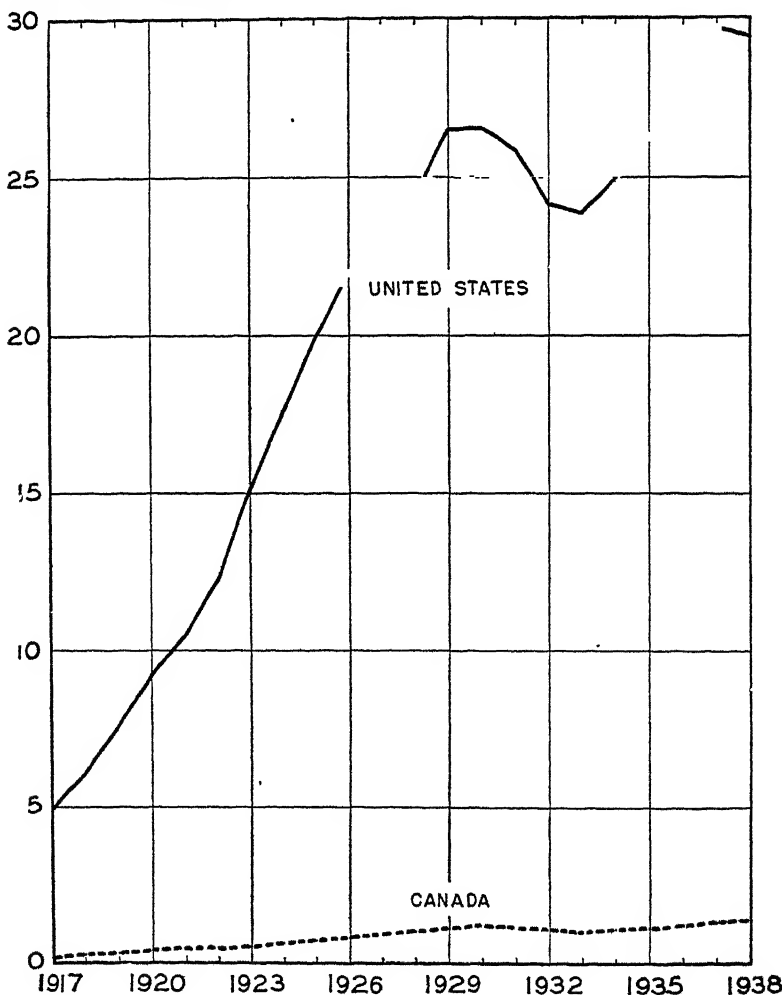
Chart 36.   Motor Vehicle Registrations in the United States and Canada, 1917-1938. (Data from Automobile Manufacturers Association; 1938 registration for Canada is an estimate.)

a sort of grid which will make a constant rate of increase appear as a straight line. In Table 19 the geometric progression of Table 17 and Chart 34 is again shown and with it are given the logarithms of the various numbers. Examination of these logarithms reveals that they form an arithmetic progression; therefore, if these logarithms are plotted on an arithmetic grid, a straight line will result, as may be seen in Chart 37. This is

TABLE 19

A GEOMETRIC PROGRESSION AND LOGARITHMS OF THE
GEOMETRIC PROGRESSION

Year (X value)	Y value	Logarithm of Y value	Amount of increase of logarithms
1931	128	2.107210	
1932	192	2.283301	.176091
1933	288	2 459392	.176091
1934	432	2 635484	.176092*
1935	648	2.811575	.176091
1936	972	2 987666	176091
1937	1,458	3 163758	.176092*
1938	2,187	3.339849	.176091

* These values differ slightly because logarithms have been rounded to nearest millionth
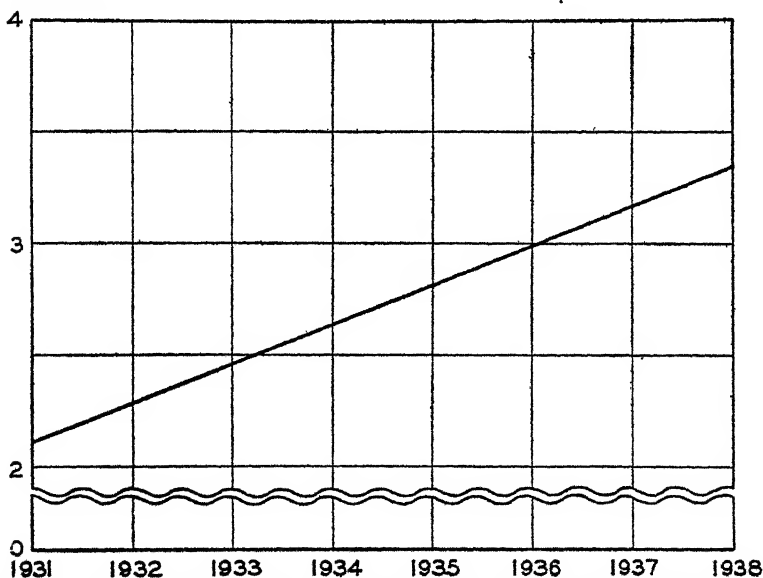
Chart 37. Logarithms of a Geometric Progression Plotted on an Arithmetic Grid. (Data of Table 19.)

one way of accomplishing our objective, but it involves the additional step of looking up logarithms before the data can be plotted. However, instead of plotting the logarithms of the values of a series, we use a grid which is designed with a logarithmic vertical scale, as in Chart 38. Here, again, we find that the geometric progression appears as a straight line. A grid of this type is termed *semi-logarithmic* because one scale is logarithmic and the other is arithmetic. Because of the assistance which this type of ruling
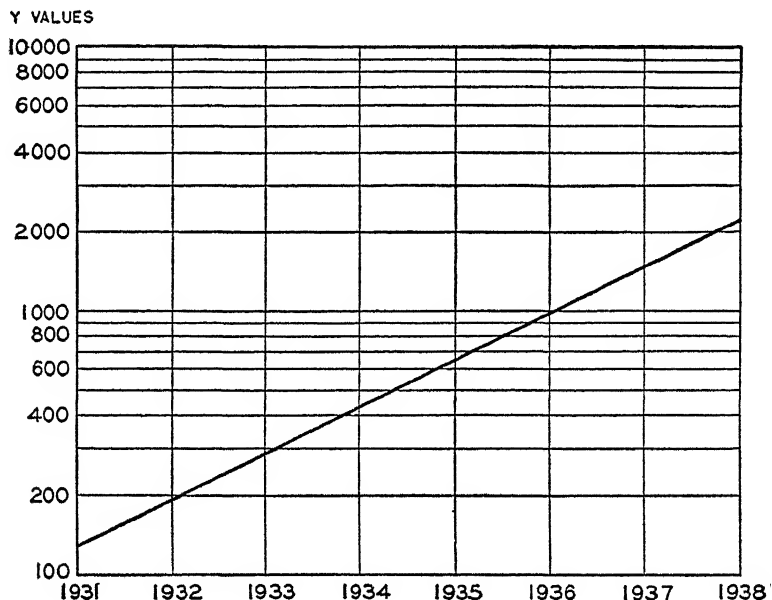
Chart 38. A Geometric Progression Plotted on a Semi-Logarithmic or Ratio Grid. Printed semi-logarithmic forms have intermediate rulings more closely spaced than those shown in this chart These closely spaced lines are an aid to plotting but are omitted from most of the charts in this book since reduction to fit the size of the page results in bringing these lines very close together. The detailed ruling is shown in Chart 52. (Data of Table 17.)

renders in comparing rates and ratios, it is frequently called *ratio* ruling.

How it is made. The construction of the logarithmic scale merely involves spacing the vertical scale values in proportion to the differences between their logarithms. Referring to Chart 39, it will be found that the distance from 2 to 3 on the scale is .352 inch, and from 3 to 4 is .250 inch. We then have:

$$\frac{\log 3 - \log 2}{\log 4 - \log 3} = \frac{.352 \text{ inch}}{.250 \text{ inch}},$$

$$\frac{.477 - .301}{.602 - .477} = \frac{.352 \text{ inch}}{.250 \text{ inch}},$$

and the proportion is:

$$.176 : .125 : : .352 \text{ inch} : .250 \text{ inch}.$$

An alternative approach to an understanding of the logarithmic scale does not involve any reference to logarithms. Reference to Chart 34 will recall that equal distances on the vertical scale of an arithmetic grid represent equal *amounts*. Equal distances measured along a logarithmic scale, however, represent equal *ratios*. On the vertical scale of Chart 38 it may be seen that the distance from 100 to 200 is .42 inch; likewise the distance from 300 to 600 is .42 inch. Measurement will reveal that any two numbers of ratio 1 : 2 are separated by .42 inch on this scale. On this same scale the distance from 200 to 800 is 84 inch, and it follows that any two numbers of ratio 1 : 4 will be separated by .84 inch. Thus we see why the semi-logarithmic chart is frequently termed the *ratio chart*.

The logarithmic scale. The vertical scale of Chart 38 is divided into two parts which are generally referred to as *cycles* or *phases*. We therefore refer to the paper on which Chart 38 was drawn as "two-cycle (or two-phase) semi-logarithmic paper." In labeling the vertical scale of a semi-logarithmic chart, we may begin with any positive value. The figure at the top of the first cycle will be ten times that at the bottom of the cycle; the figure at the top of the second cycle will be ten times the figure at the bottom of the second cycle

Chart 39. The Logarithmic Scale. The vertical distances are proportional to the differences between the logarithms. Each vertical distance is twice the difference between the logarithms measured in inches.

(the top of the first cycle); and so on.[3] In Chart 40 there are illustrated eight different logarithmic scales beginning with .1, 1, 2, 5, 10, 17, 25, and 50 respectively. Although it is mathematically permissible to begin a logarithmic scale with any positive value, it is advisable to select a scale which will allow interpolations of intermediate values to be made readily. The scale beginning with 17 would be very difficult to use. If it were desired to have a three-cycle scale beginning with .5, the various values

[3] A common logarithm is the power to which 10 must be raised to produce a given number. Thus, 100 is 10^2, and the logarithm of 100 is 2.0; 10,000 is 10^4, and the logarithm of 10,000 is 4.0.

of the first scale could be multiplied by 5. Most ready-ruled semi-logarithmic paper carries along the right-hand edge of the grid such
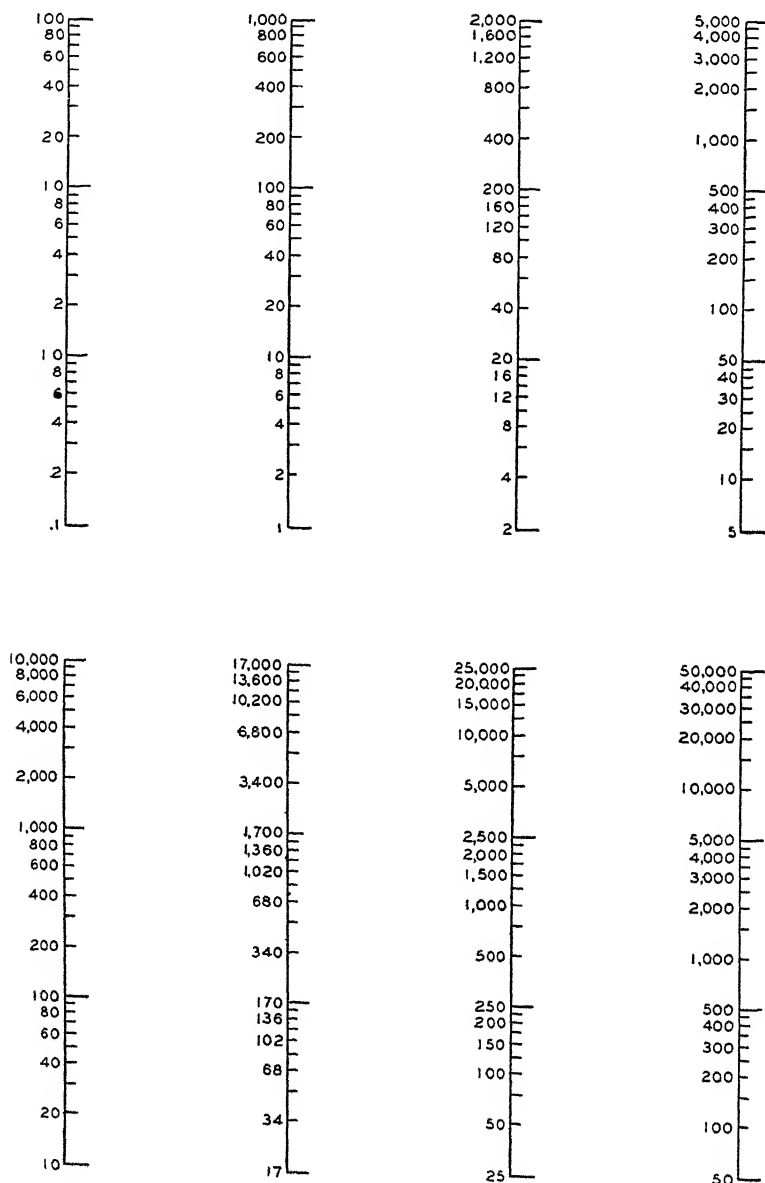
Chart 40. Logarithmic Vertical Scales. The scale beginning with 17 would be difficult to use.

designations as those shown in Chart 52. These are multiplying factors and indicate that the value to be written opposite each horizontal line on the left-hand scale must be the value at the bottom of that cycle multiplied by the figure shown opposite that horizontal line on the right-hand scale.

If a logarithmic scale were begun with zero, the top of the first cycle would be $10 \times 0 = 0$, and all values on the scale would also be zero. Suppose that the uppermost value of a three-cycle logarithmic scale is .01. Then the bottom of the third cycle is $\frac{1}{10}$ of .01, or .001; the bottom of the second cycle is .0001; and the bottom of the first cycle is .00001. There can thus be no zero base line, and the semi-logarithmic chart does not permit interpietation of curves in terms of distances above a base line as does the arithmetic chart. Although plotted values may, of course, be read against the vertical logarithmic scale, no visual impression may be had of the absolute magnitudes plotted. The semi-logarithmic chart shows: (1) a constant rate of change as a straight line; (2) the rate of increase or decrease by the slope of the line; and (3) the comparison of rates or ratios between two or more lines by means of parallelism of these lines, or lack of it.

Interpretation of curves. Before proceeding with a consideration of applications of the semi-logarithmic chart, we should give attention to Charts 41A and 41B and the comments below them. When two curves are parallel on semi-logarithmic paper (for example, a, a'; d, d'), we know that they have undergone the same rate of change and also that the ratio between the two has remained constant. Parallelism between curved lines is very difficult to judge with the eye. Reference to the lower sections of Chart 41A will show that the curved lines are always the same *vertical* distance apart and thus the two curves in each section are parallel with respect to the X axis.

Applications

Comparing rates of increase or decrease. Since there is no zero on the vertical scale of the semi-logarithmic chart and thus no base line, and since equal vertical distances (on the same scale) always represent the same ratio, it is permissible to use two or more different vertical scales in order to bring curves of different magnitude close together for comparison. This has been done in Chart 42 which presents the data of motor vehicle registrations previously shown on an arithmetic grid in Chart 36. Shifting the vertical scale of a semi-logarithmic chart moves the curve upward or downward, but the slope, which is of paramount importance, is not altered thereby. When using two logarithmic scales, as in Chart 42, it is highly desirable (though not absolutely necessary) to keep the series of smaller
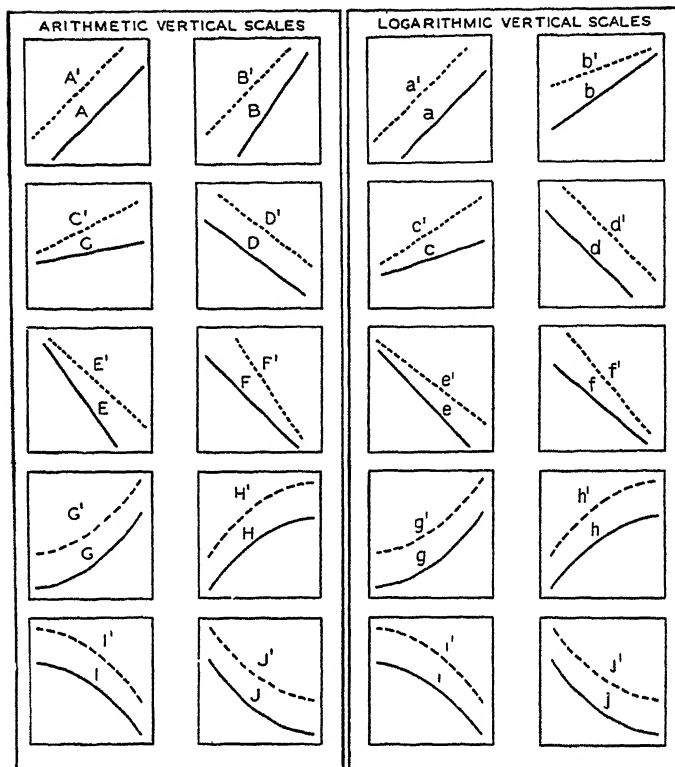
Chart 41A.   Curves on Arithmetic and Semi-Logarithmic Grids.   The two curves in each of the lower eight squares are equidistant vertically from each other.

ARITHMETIC VERTICAL SCALES

A, A'—Constant amounts of increase, same for both curves
B, B'—Different constant amounts of increase, greater for B
C, C'—Different constant amounts of increase, greater for C'.
D, D'—Constant amounts of decrease, same for both curves
E, E'—Different constant amounts of decrease, greater for E.
F, F'—Different constant amounts of decrease, greater for F'.
G, G'—Amounts of increase increasing, same for both curves.
H, H'—Amounts of increase decreasing, same for both curves.
I, I'—Amounts of decrease increasing, same for both curves.
J, J'—Amounts of decrease decreasing, same for both curves.

LOGARITHMIC VERTICAL SCALES

a, a'—Constant rates of increase, same for both curves
b, b'—Different constant rates of increase, greater for b.
c, c'—Different constant rates of increase, greater for c'.
d, d'—Constant rates of decrease, same for both curves.
e, e'—Different constant rates of decrease, greater for e
f, f'—Different constant rates of decrease, greater for f'.
g, g'—Rates of increase increasing, same for both curves.
h, h'—Rates of increase decreasing, same for both curves.
i, i'—Rates of decrease increasing, same for both curves.
j, j'—Rates of decrease decreasing, same for both curves.

ARITHMETIC VERTICAL SCALES
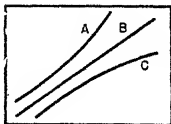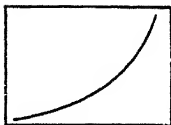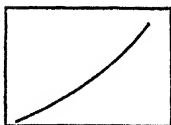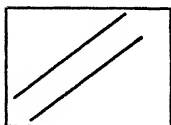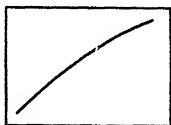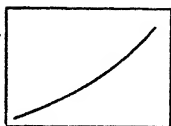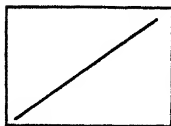
LOGARITHMIC VERTICAL SCALES
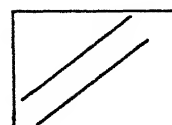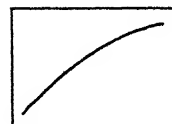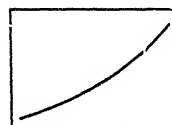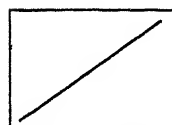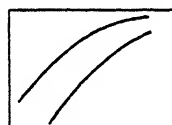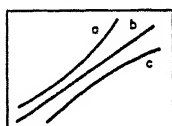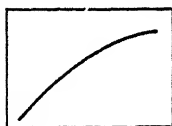
An arithmetic progression.

A series in which the absolute change is increasing
- a   If relative change is increasing.
- b   If relative change is constant
- c   If relative change is decreasing

A series in which the absolute change is decreasing

Two arithmetic progressions, same absolute changes.

A geometric progression.

A series in which the relative change is increasing

A series in which the relative change is decreasing
- A   If absolute change is increasing
- B.   If absolute change is constant
- C   If absolute change is decreasing

Two geometric progressions, same relative changes

Chart 41B.   Comparisons of Series of Various Types Plotted in Relation to Arithmetic and Logarithmic Vertical Scales.   Series plotted as shown on one scale become as indicated on the other.   The above comparisons refer to increasing series only.   It is suggested that the reader sketch some comparisons involving declining series.

111

magnitude below that of greater magnitude; likewise, if one or more components are being compared with a total, the curves for the components should be below that for the total.

Chart 36 gave us no idea of the *relative* growth of automobile registrations in either the United States or in Canada. Chart 42, however, shows relative growth for each series and enables us to compare the rates of growth of these two series of dissimilar size. In general both series have shown about the same rates of increase and decrease throughout the period. The

UNITED STATES
MILLIONS OF
MOTOR VEHICLES

CANADA
THOUSANDS OF
MOTOR VEHICLES
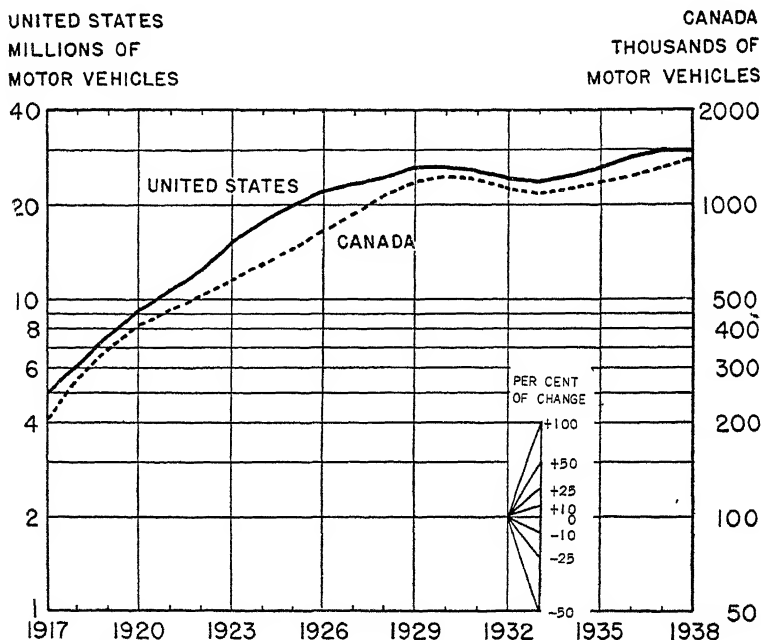
Chart 42. Motor Vehicle Registrations in the United States and Canada, 1917-1938. (Data from Automobile Manufacturers Association; 1938 registration for Canada is an estimate)

insert on Chart 42 makes it possible to estimate the rate of increase or decrease for the curves shown. It does not, however, apply to other charts which have different scales.

An alternate method of showing the relative change in motor vehicle registrations in the United States and Canada consists of calculating the per cent of change for each year and plotting the results on an arithmetic grid. This has been done in Chart 43.

Instead of comparing the rates of change of two different series over the same period of time, we may be interested in comparing rates of growth of

the same series at different times.   Thus in Chart 42 we can see that the rate of increase of United States automobile registrations was greater from 1935 to 1936 than from 1936 to 1937, and also that the rate of decline was greater from 1931 to 1932 than from 1932 to 1933.   Similar conclusions may be drawn from Chart 43.

It is frequently necessary to compare series which are expressed in different units.   For example, we may compare any two or more of the following: commercial failures, in millions of dollars; volume of trading on a stock exchange, in number of shares traded; coal production, in 2,000-pound tons; petroleum production, in 42-gallon barrels; lumber production,
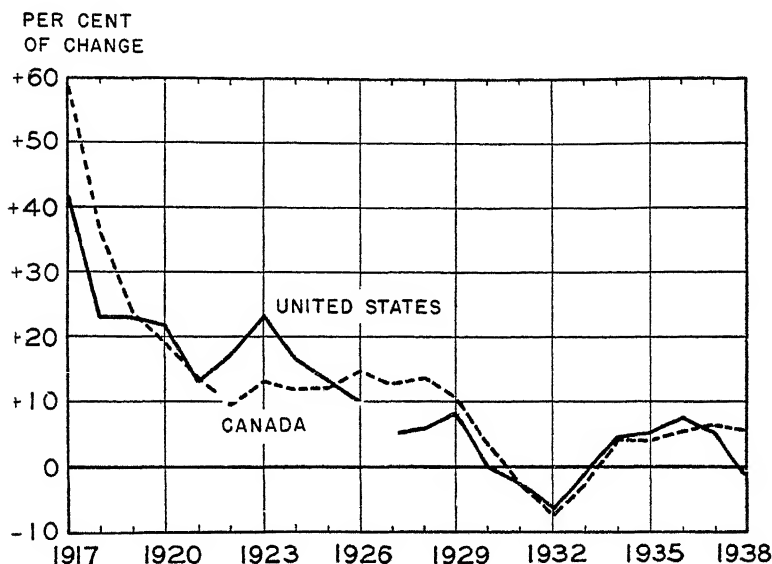
PER CENT
OF CHANGE

Chart 43.   Annual Per Cent of Increase or Decrease in Motor Registrations in the United States and in Canada, 1917-1938.   (Data from same source as Chart 42.)

in board feet; cement production, in 350-pound barrels; electric power produced, in kilowatt hours; manufactured gas, in cubic feet.   It is possible to reduce 350-pound barrels to tons, but it is not possible to change kilowatt hours to board feet, or vice versa.

While it is possible to plot two series expressed in different units on an arithmetic grid, it is not often that such a comparison is useful.   We are not likely to be interested in comparing the changes in electric power production in kilowatt hours with the changes in cement production in barrels. Rather are we apt to want to compare the percentage change in electric power production with the percentage change in cement production.   On the semi-logarithmic grid there is no zero base line; only the slope of a

curve has meaning, and we are enabled to make a valid comparison of the relative changes in the two series expressed in such dissimilar units as those just mentioned.   Chart 44 shows a comparison of the production of electric energy and of portland cement.   Among other interesting comparisons may be noted the more rapid relative growth in the production of electric energy from 1921 to 1929 and the relatively more severe decline in production of cement from 1929 to 1933.

Comparing fluctuations.   The comparison of the fluctuations taking place in two series of different size may be illustrated by reference to the
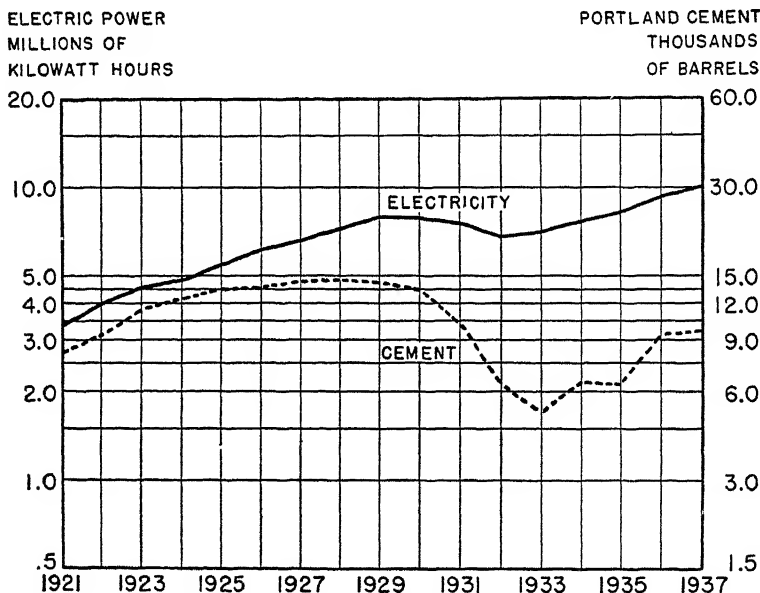
Chart 44.   Average Monthly Production of Electric Power and of Portland Cement, 1921-1937.   (Data from *Survey of Current Business*, 1938 Supplement, p. 150)

1902–1938 annual prices of pig iron and of finished steel, the latter being an average of quotations on steel bars, beams, tank plates, plain wire, open-hearth rails, black pipe, and black sheets.   The quotations for pig iron are given in dollars per long ton and varied from a low of $12.95 in 1914 to a high of $42.76 in 1920.   Finished steel, on the other hand, is quoted in terms of cents per pound, and the price varied from 1.433 cents in 1914 to 4.191 cents in 1917.   If a vertical scale were designed to accommodate the quotations for pig iron, it is easy to see that the steel price curve would virtually coincide with the base line.   If a vertical scale were designed to fit the prices of steel, the curve for pig iron prices would be above the top

of any reasonable-sized piece of graph paper. Using a single arithmetic vertical scale, we cannot plot these two series on the same axes.

However, we can transform the prices of finished steel to prices per ton. The results are shown in Chart 45. Here it is seen that the prices of finished steel are higher than the prices of pig iron. We can also note the highs and lows for each series, and can observe that the *absolute* fluctuation in prices during 1914–1924 was less for pig iron. If we are interested in the severity of the *relative* fluctuations, we should examine the curves of Chart 46. This semi-logarithmic chart shows clearly that the relative fluctua-
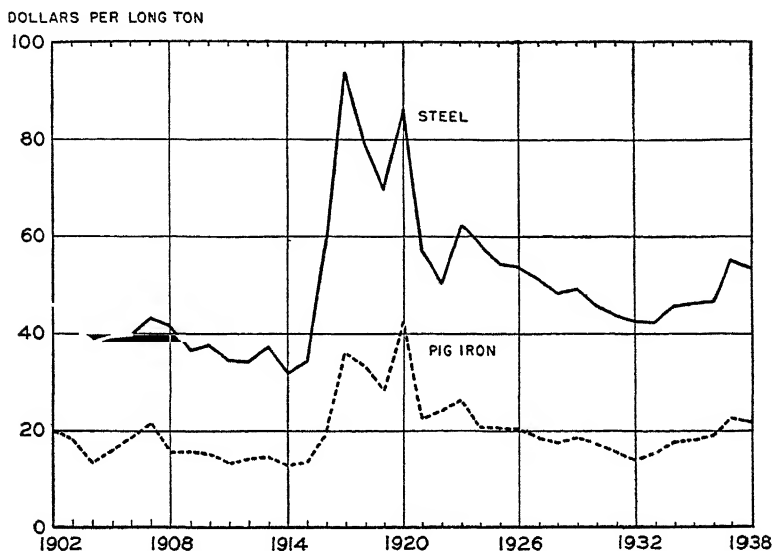
Chart 45. Price per Long Ton of Pig Iron and of Finished Steel. (Data from Standard Statistics Company, *Basic Statistics*, 1936, pp. G6 and G12, and various issues of *Current Statistics*. Both price series are composites.)

tions during 1914–1924 were about the same. Note that, while the absolute rise from 1919 to 1920 was greater for steel (Chart 45), the relative increase was greater for pig iron (Chart 46). Note also that a greater relative rise is shown for pig iron during the period 1932–1934.

In order to plot these two series on the arithmetic chart, it was necessary to express steel prices on a per ton basis. Such an adjustment would not have been possible if we had been dealing with one series expressed in terms of tons and another in terms of yards (say, for example, iron and rayon). Although we used both price series in terms of dollars per ton for Chart 46, the semi-logarithmic chart does not impose such a require-ment. When we are interested only in the relative change which has

taken place, it does not matter if either the money unit or the physical unit or both differ for the two series.   Chart 47 shows the relative fluctuations in the prices of pig iron (in dollars per ton) and of steel (in cents per
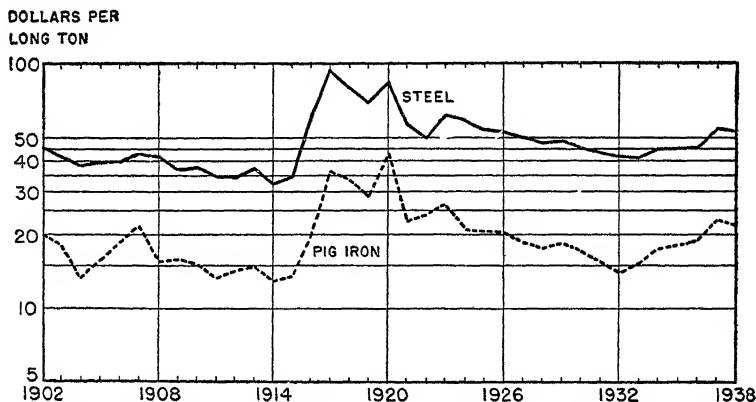
Chart 46.   Price per Long Ton of Pig Iron and Finished Steel, 1902–1938.   (Data from same source as Chart 45)

pound).   Notice that the steel price curves of Charts 46 and 47, although the prices are in terms of different units, would coincide if superimposed upon each other.
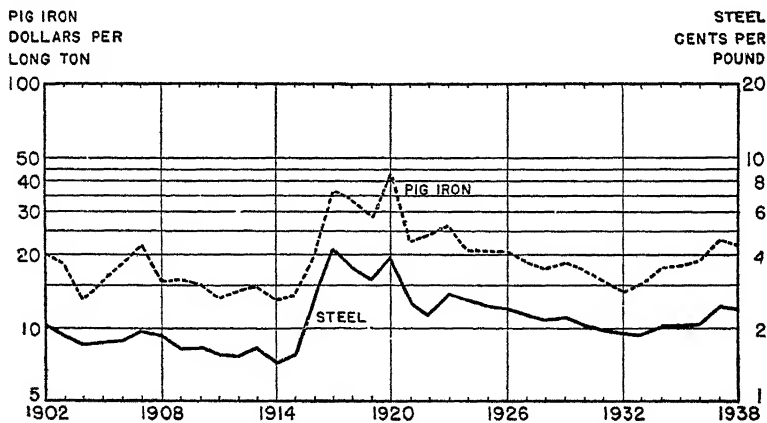
Chart 47.   Price of Pig Iron per Long Ton and of Finished Steel per Pound, 1902–1938. (Data from same source as Chart 45)

Instead of being interested in two series, we may wish to compare the undulations of a single series which fluctuated around relatively small values during one period and around decidedly larger values at another time.   For example, commercial failures were around $100,000,000 to

$200,000,000 from 1895 to 1910.   From 1921 to 1933, however, they ranged from $400,000,000 to $933,000,000.   The semi-logarithmic chart enables us to study the relative severity of the fluctuations during such different periods.

Showing ratios.   Chart 48 shows how ratios may be presented on the semi-logarithmic chart.   The two series plotted are the price per bushel received by farmers for corn, and the price per 100 pounds received by farmers for hogs.   When corn is bringing a price which is low in relation
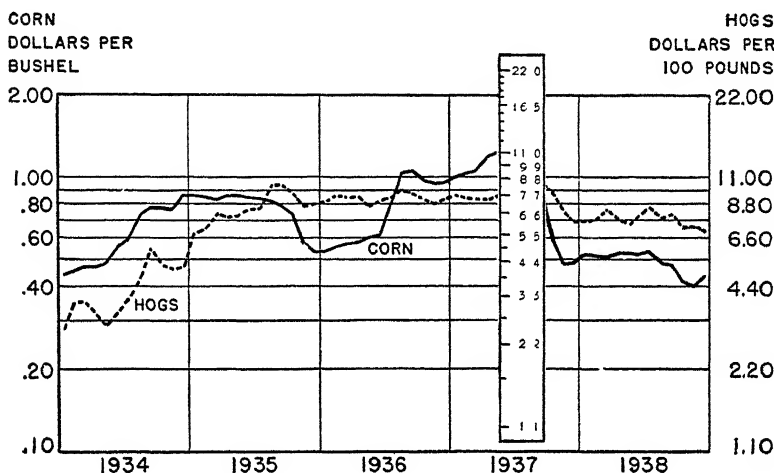
Chart 48. Average Farm Prices of Corn per Bushel and of Hogs per Hundred Pounds, 1934-1938. The supplementary scale enables us to read the ratio of hog prices to corn prices for any month. The value 11 is placed opposite the line for corn, and the value opposite the hog line is the ratio of hog prices per 100 pounds to corn prices per bushel. For May 1937 the ratio is shown to be 7.7, which may be verified by referring to Chart 49. The supplementary scale is graduated the same as the scales at the sides of the chart. The figure 11 is placed opposite the corn line because the hog scale of the diagram shows values which are 11 times the corresponding values on the corn scale. (Data from various issues of *Crops and Markets*.)

to the price of hogs, farmers will generally find it profitable to feed corn to hogs rather than to sell the corn for cash.   On the other hand, when corn is bringing a price which is high in relation to that of hogs, farmers will tend to sell corn for cash.   If 100 pounds of hogs brings the farmer about 11 times as much as a bushel of corn, it is largely immaterial to the farmer whether he sells his corn for cash or feeds the corn to his hogs.[4]   For this reason the two scales of Chart 48 have been placed in an 11 to 1 ratio.[5]   The chart shows not only the fluctuations in the price of hogs and the price

---

[4] See pp. 155–156, where the hog-corn ratio is discussed.
[5] The scale for hog prices is awkward but is unavoidable in this instance.

of corn, but also makes it easy to see when the price of 100 pounds of hogs is more than, less than, or exactly, 11 times the price of a bushel of corn. When 100 pounds of hogs is selling for more than 11 times as much as a bushel of corn, the curve for hogs is above the curve for corn, hogs are relatively valuable, and farmers tend to feed corn to their hogs.   When 100 pounds of hogs is selling for less than 11 times as much as a bushel of corn, the curve for hogs is below that for corn, corn is relatively valuable, and farmers tend to sell corn for cash.   When the two curves are parallel, the ratio is remaining constant; when the corn price curve is sloping upward *more* rapidly (or downward *less* rapidly) than the hog price curve, corn is
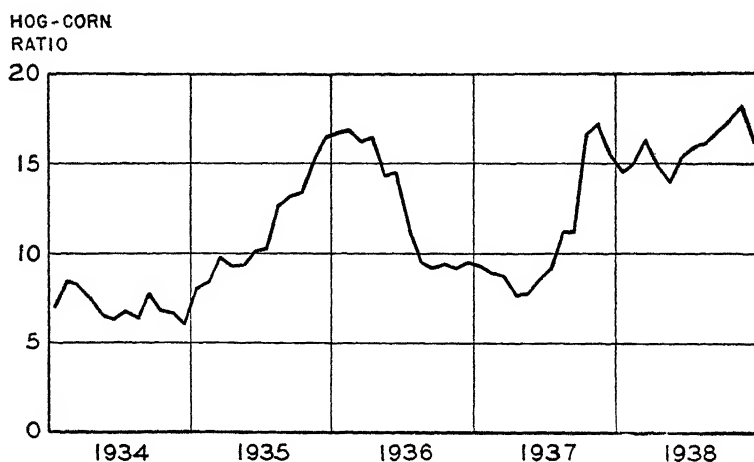
Chart 49.   Hog-Corn Ratio, 1934-1938.   The ratio is obtained by dividing the average farm price of hogs per 100 pounds by the average farm price of corn per bushel. (Data from *Crops and Markets*, January 1939, p. 10)

becoming *more* valuable in relation to hogs; when the corn price curve is sloping upward *less* rapidly (or downward *more* rapidly) than the hog price curve, corn is becoming *less* valuable in relation to hogs.   The supplementary scale, shown on the chart, enables the reader to measure the ratio between the two price curves at any time.

Chart 49 illustrates another method of showing the relationship between hog and corn prices.   Here the ratio of hog prices to corn prices has been computed for each month and plotted on an arithmetic grid.   The ratio may be studied without the use of a supplementary scale, but changes in corn prices and in hog prices are not shown.

Interpolation and extrapolation.   While an interpolation on an arithmetic chart is an arithmetic interpolation, an interpolation on the semilogarithmic chart is a logarithmic interpolation.   Thus, if we refer to

Chart 38 and graphically interpolate for the Y value midway between 1935 and 1936, we obtain about 790, which is approximately the same figure that we get if we use (log 648 + log 972) ÷ 2 and take the anti-logarithm of the result.

Extrapolation consists of extending the curve at one end or the other. When we extend the curve to estimate for later years than those for which we have data, we are forecasting. This application of the semi-logarithmic chart is decidedly of questionable value if it involves merely the extension of a curve which has indicated in the past that the data exhibit a fairly constant rate of increase. Any forecasting procedure which involves
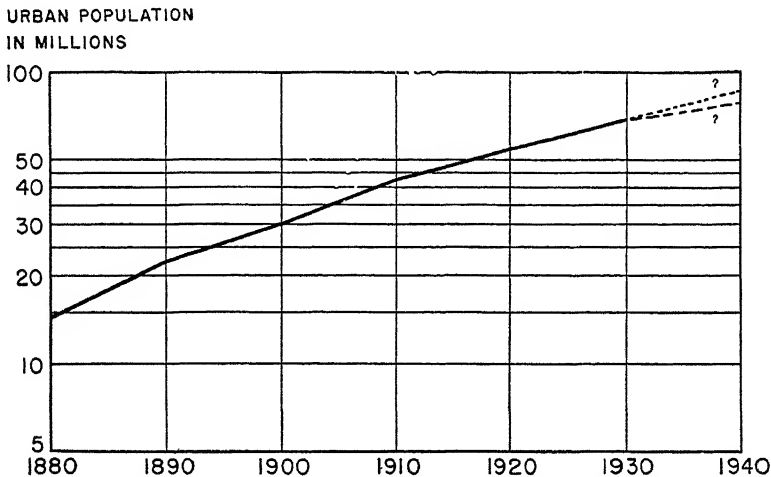
Chart 50. Urban Population of the United States, 1880-1930, and Two Estimates for 1940. A dubious application of the semi-logarithmic chart. (Data from *Fifteenth Census of the United States, 1930*. Population Volume I, p. 8. A slight change in the classifications "urban" and "rural" took place in 1930. See p 46 of this text.)

merely the continuation of a curve or the automatic application of a formula, without at the same time requiring a careful consideration of underlying and modifying factors, is hardly to be depended upon, particularly if economic conditions are in a state of flux. The curve of Chart 50 shows the population of the United States classed as "urban" at each census from 1880 to 1930 inclusive. While the extensions of the curve indicate two possible estimates for 1940, it should be realized that any estimate of urban population in 1940 based *only* on a knowledge of the six preceding censuses can have little validity. What of the subsistence farming movement? What of decreased immigration? What of birth control? What of those who went back to live on the home farm in the depression years?

Flexible Logarithmic Scales

One logarithmic cycle will accommodate a ten-fold increase; two cycles make provision for a hundred-fold increase. Reference to the various charts included in this chapter will show that no vertical logarithmic scale extends over more than two cycles. Two-cycle semi-logarithmic paper will suffice for most series which the chart maker is likely to encounter; rarely will he need paper covering more than three cycles, since it allows for a thousand-fold increase. Even in cases where a series of very small magnitude must be compared with one of very large magnitude, a number of cycles is not needed, since it is desirable to use two vertical scales to bring the two curves together for comparison, as in Chart 42. Many sorts of ready-ruled semi-logarithmic paper are available from various sources. If, however, only two-cycle paper is available and paper having more cycles is needed, it is merely necessary to trim the lower margin from a sheet of two-cycle paper and paste it above another sheet.

At times it may be desirable to use one- or two-cycle paper, but with a larger or smaller size cycle than those which are readily available. Using an ordinary sheet of semi-logarithmic paper and placing a sheet of plain paper diagonally on top of it, a logarithmic scale may be expanded as shown in Chart 51. A logarithmic scale may be contracted by placing a sheet of semi-logarithmic paper diagonally on a piece of plain paper and ruling horizontal lines as shown in Chart 52. For those who have frequent occasion to use logarithmic scales of varying size, a device such as that shown in Chart 53 is useful.[6] The original of this chart provides a logarithmic

Scale value	Logarithm	Difference
1	0	. .
2	.301030	.301030
3	.477121	.176091
4	.602060	.124939
5	.698970	.096910
6	.778151	.079181
7	.845098	.066947
8	.903090	.057992
9	.954243	.051153
10	1.000000	.045757
20	1.301030	.301030
30	1.477121	.176091
40	1.602060	.124939
50	1.698970	096910
60	1.778151	.079181
70	1.845098	.066947
80	1.903090	.057992
90	1.954243	.051153
100	2.000000	.045757

---

[6] Purchasable from Harriet Edmunds, 202 East 44th Street, New York City.
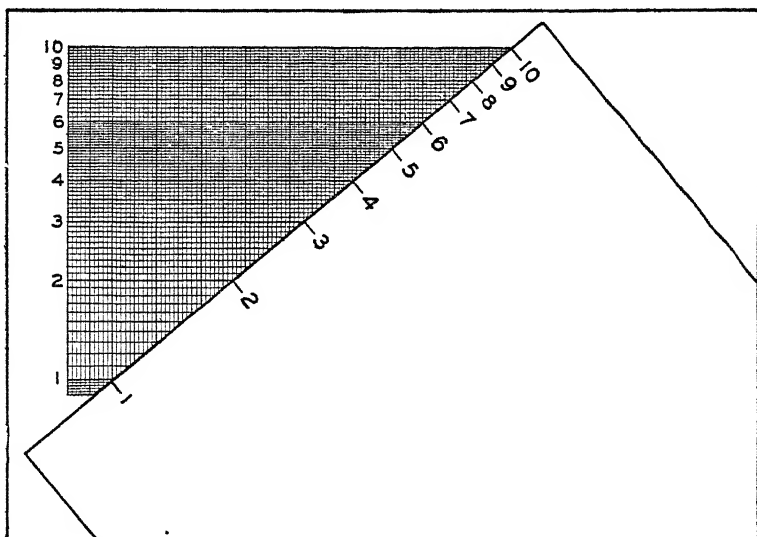
Chart 51.   A Method of Expanding a Logarithmic Scale.
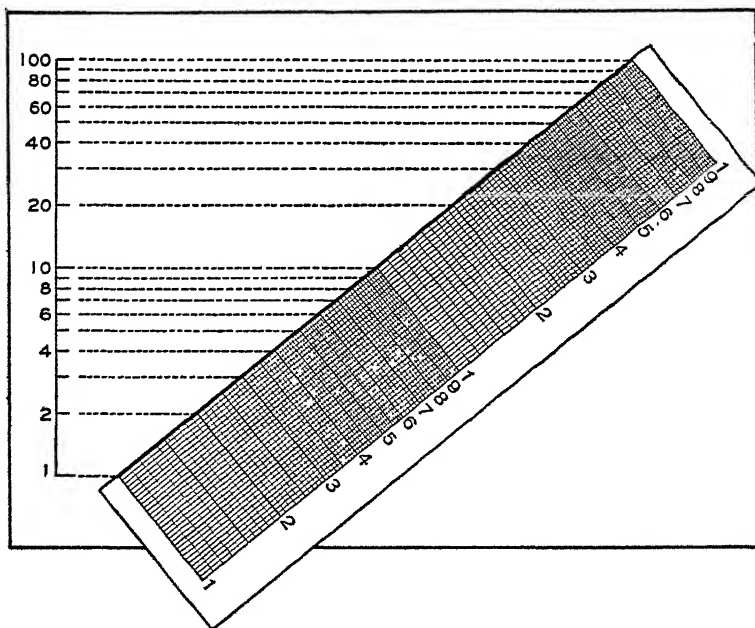
Chart 52.   A Method of Contracting a Logarithmic Scale.
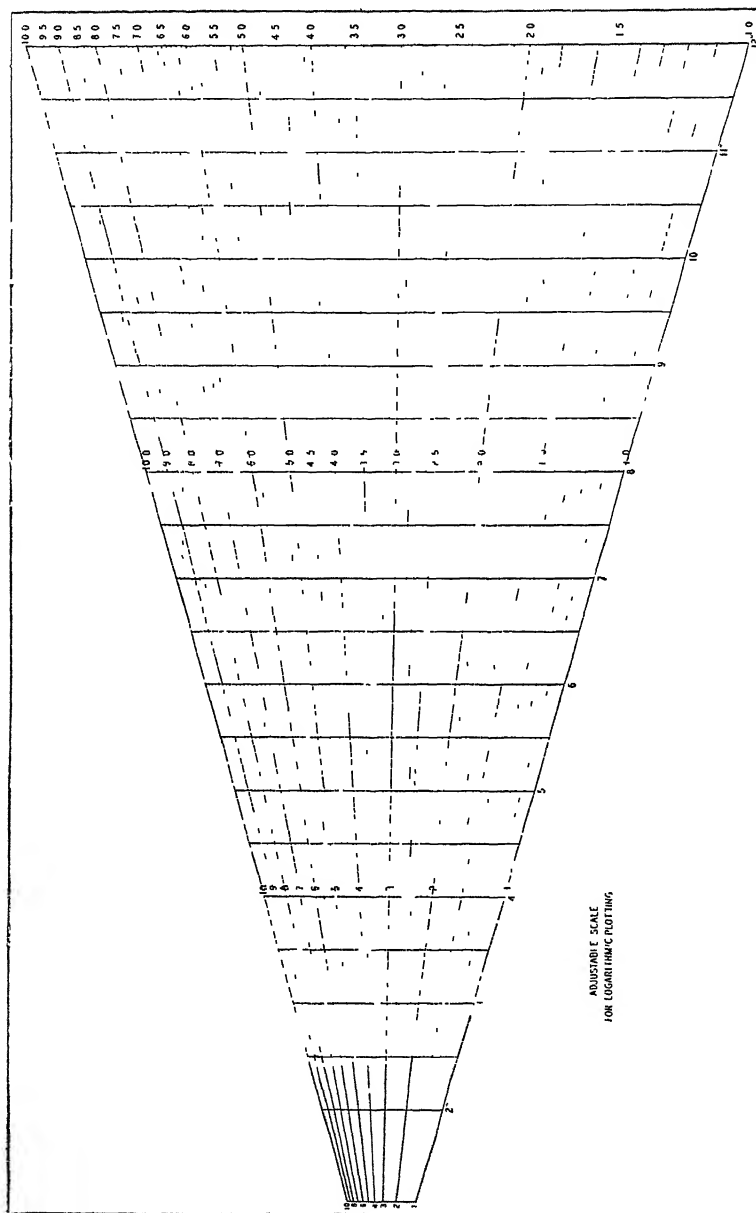
Chart 53.   A Flexible Logarithmic Scale.   The original provides logarithmic scales ranging from 1⅓ to 12 inches

cycle varying from $1\frac{1}{8}$ inches to 12 inches. Of course, any number of cycles may be built up on top of one another.

In case no suitable logarithmic paper and no logarithmic scales of any sort are available, it is possible to construct a logarithmic scale of any desired size by referring to a table of logarithms. With scale values spaced in proportion to the differences between their logarithms, a scale may be constructed in terms of any convenient unit. From the figures on page 120 it is seen that the distance from 1 to 2 would be .301030 units, the distance from 2 to 3 would be .176091 units, and so on. Intermediate values·are located similarly.

The usefulness of logarithmic scales is not limited to the applications shown in this chapter. In Chapter XI we will make use of a horizontal logarithmic scale and an arithmetic vertical scale. In Chapters VIII and XXIII we will use logarithmic scales on both the horizontal and vertical axes.

Selected References

P. A. Bivins: *The Ratio Chart in Business,* Codex Book Co., Norwood, Mass., 1926. A comprehensive treatment.

W. C. Brinton· *Graphic Presentation,* Chapter 41; Brinton Associates, 608 West 45th St., New York, 1939.

L. W. Chaney: "Comparison of 'Arithmetic' and 'Ratio' Charts," *Monthly Labor Review,* Vol. VIII, No. 3, March 1919, pages 20–34.

Irving Fisher: "The 'Ratio' Chart for Plotting Statistics," *Quarterly Publication of the American Statistical Association,* Vol. 15, No 117, June 1917, pages 577–601. An excellent exposition.

A. C. Haskell: *Graphic Charts in Business* (Second Edition), Chapters V–VII, X, XI; Codex Book Co., Norwood, Mass., 1928.

K. G. Karsten: *Charts and Graphs,* Chapter XXXVI; Prentice-Hall, Inc., New York, 1923.

J. R. Riggleman *Graphic Methods for Presenting Business Statistics* (Second Edition), Chapter V; McGraw-Hill Book Co., New York, 1936.

CHAPTER VI

GRAPHIC PRESENTATION
OTHER TYPES OF CHARTS

Not only may we use curves to present statistical information, but there are available a number of other graphic devices as well. In this chapter we shall give attention to bar charts, pie diagrams, pictorial charts, and statistical maps.

Bases of Comparison

Chart 54 shows how the number of tractors on farms may be compared by means of three types of diagrams: (A) a bar chart involving one-dimensional comparisons; (B) and (C) circles and squares, involving two-dimensional comparisons; and (D) a three-dimensional comparison represented by tractors of varying sizes. Readers of charts obtain most accurate impressions of the magnitudes shown when data are represented by means of bar charts, and least accurate impressions when data are represented by volume diagrams. Area diagrams are more accurately judged than volume diagrams, but less accurately than bar charts.[1] It should also be remembered that volume diagrams shown on the printed page make it necessary for the reader to visualize the third dimension before making his comparison. Another disadvantage of charts using squares, circles, or pictures of different sizes is that the reader may be uncertain whether to compare heights, areas, or volumes. In any event the basis upon which the diagram was drawn should be indicated. If it is argued that the correct basis of comparing the size of such objects as tractors is the apparent weight of the different tractors, but if the statistician has drawn the tractors so that the number of tractors in different years is shown by the height of the tractors, as is often done, then the reader who judges the sizes upon the basis of apparent weight (essentially volume) will get an exaggerated impression of the variation in number of tractors during the different years.

---

[1] See "Graphic Comparisons by Bars, Squares, Circles, and Cubes," by Frederick E. Croxton and Harold Stein, *Journal of the American Statistical Association*, March 1932, pp. 54–60.

A

B

C

D
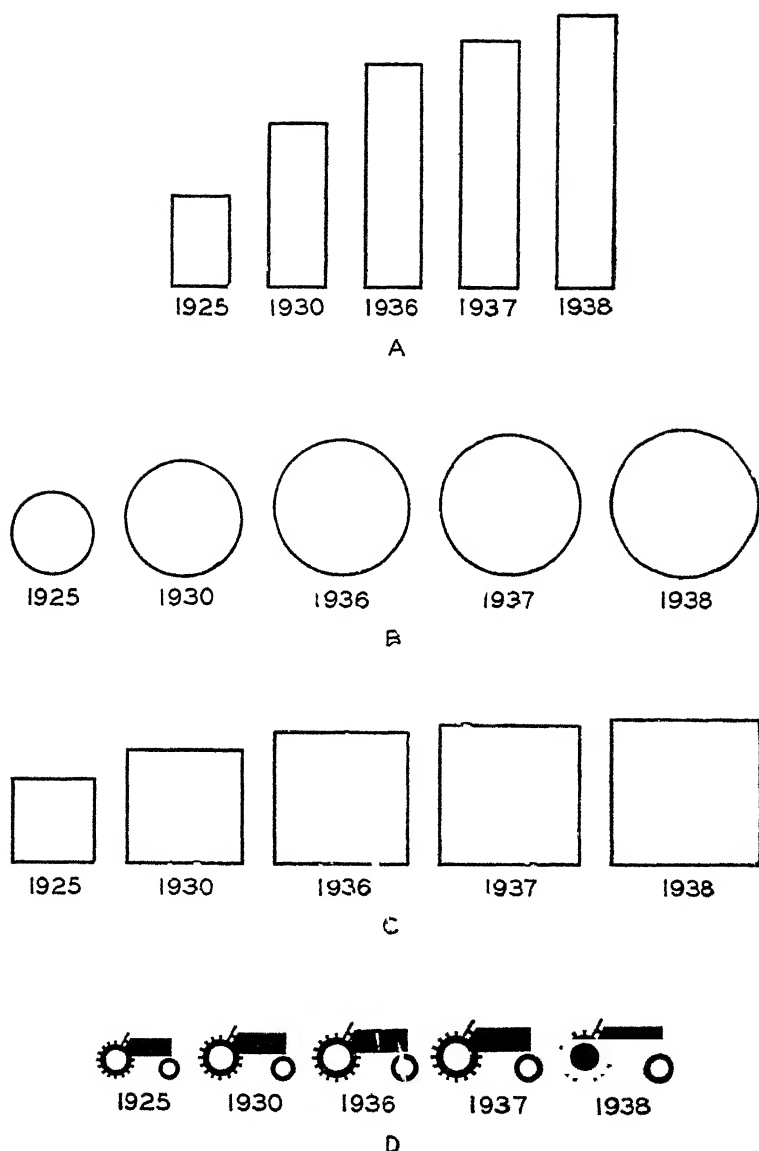
Chart 54. Number of Tractors on Farms in the United States, 1925, 1930, and 1936-1938. Represented by (A) bars, (B) circles, (C, squares, and (D) tractors. Parts B and C show the comparisons by areas, while part D shows the comparisons by volumes. (Data for 1925 and 1930 from *Fifteenth Census of the United States, 1930*, Agriculture Volume II, Part I, p. 55. Data for 1936-1938 are from *Farm Implement News* April 7, 1938.)

Bar Charts

The bar chart shown in section A of Chart 54 is a simplified form using no scale.   In Chart 55 the same data are shown by means of a bar chart which has a scale and which also varies the spacing between the bars in order to call attention to the fact that the time intervals vary.   When the
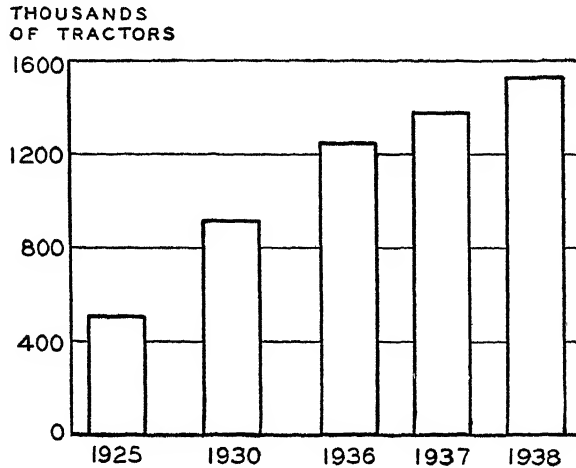
Chart 55.   Number of Tractors on Farms in the United States, 1925, 1930, and 1936-1938.   (Data from same sources as Chart 54)

chart is expected merely to convey a very general impression, simple bar charts may be drawn without the use of a scale as in section A of Chart 54. However, the scale should not be omitted when two or more bar charts are shown depicting different magnitudes.   Consider Chart 56, which shows
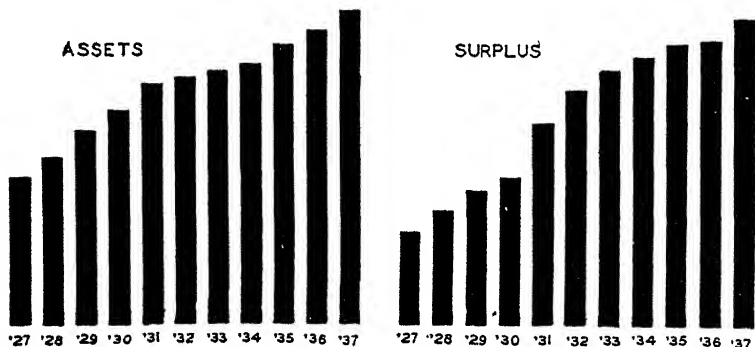
Chart 56.   Assets and Surplus of an Insurance Company, 1927-1937, as Shown in an Advertisement.   Because of the absence of vertical scales and the proximity of the two sets of bars, the chart gives the visual impression that surplus is nearly as great as assets.

the assets and surplus of an insurance company as set forth in an advertisement.   The chart shows that both assets and surplus have grown, but it also gives the impression that surplus, in the later years, was almost as great as assets.   This, of course, is incorrect.   Each part should have had a vertical scale, which would have shown 1937 assets to have been slightly over $200,000,000 and 1937 surplus to have been about $12,000,000.

All the bar charts that have been shown are representations of chronological data and, following the customary procedure, the bars have been arranged vertically.   Vertical bars should also be used for data classified

MILLIONS OF DOLLARS

| | 0 | 200 | 400 | 600 | 800 |

EUROPE $719,300,000

ASIA $707,400,000

NORTH AMERICA $619,700,000

SOUTH AMERICA $290,500,000
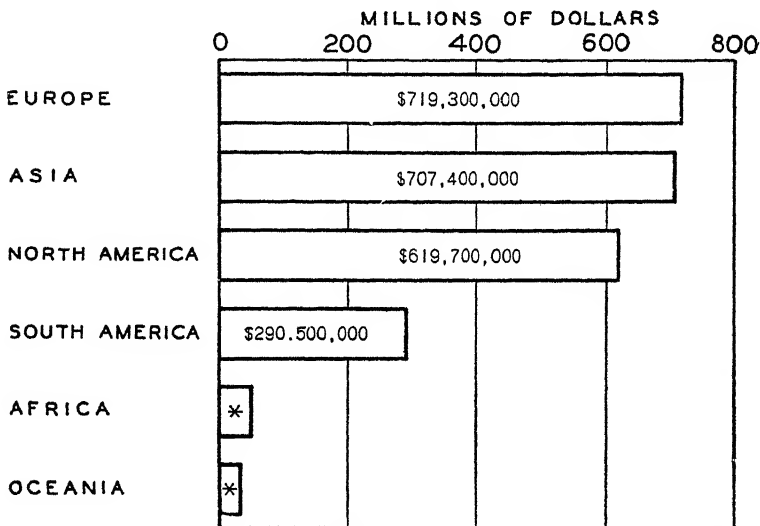
AFRICA   *

OCEANIA   *

Chart 57.   Value of United States Imports from Each Continent, 1936.   (Data from *Statistical Abstract of the United States, 1937*, p. 448)
    * Africa $51,000,000; Oceania $36,100,000.

quantitatively, as in Chart 61.   When making comparisons of data classified qualitatively or geographically, on the other hand, horizontal bars are generally used.   Chart 57 shows such a comparison of United States imports from each continent.   There are no set rules to be observed in drawing bar charts.   Certain considerations, however, are helpful.

(1) Individual bars should be neither exceedingly short and wide nor very long and narrow.

(2) Bars should be separated by spaces which are not less than about $\frac{1}{2}$ the width of a bar or greater than about the width of a bar.

(3) A scale is generally useful.   It should be not more than $\frac{1}{2}$ the width of a bar from the top bar (or from the left-hand bar if the bars are vertical).

(4) Guide lines are an aid in reading the chart.   Sometimes the chart
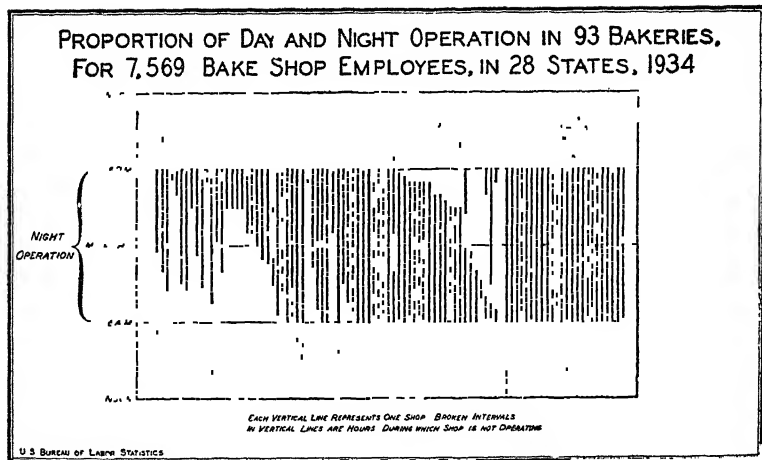
Chart 58. An Application of the Bar Chart. (From United States Bureau of Labor Statistics, *Wages, Hours, and Working Conditions in the Bread-Baking Industry, 1934*, Bulletin No. 623, p. 75)

is enclosed and the guide lines are extended through the entire chart as in Chart 57; sometimes the chart is not enclosed and the guide lines are cut off as in Chart 60.

When showing a time series graphically, we may use either a bar chart or a curve. If the series covers many years, it is generally not desirable
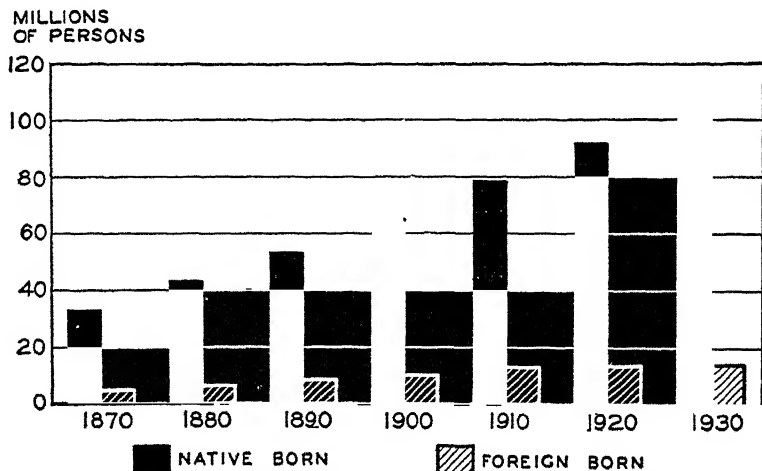
Chart 59. Native-Born and Foreign-Born Population of the United States, 1870-1930. The relative growth of the two series is not apparent from this type of chart, but may be shown by means of a semi-logarithmic chart as described in the preceding chapter. (Data from the *Statistical Abstract of the United States, 1937*, p. 11.)

to use a bar chart, which is laborious to construct. A curve facilitates a study of the general change which has taken place in a series, whereas a bar chart enables comparisons of specific years to be made more readily.

Chart 58 shows an interesting application of the principle of the bar chart. It shows for each of 93 bakeries the proportion of day and night
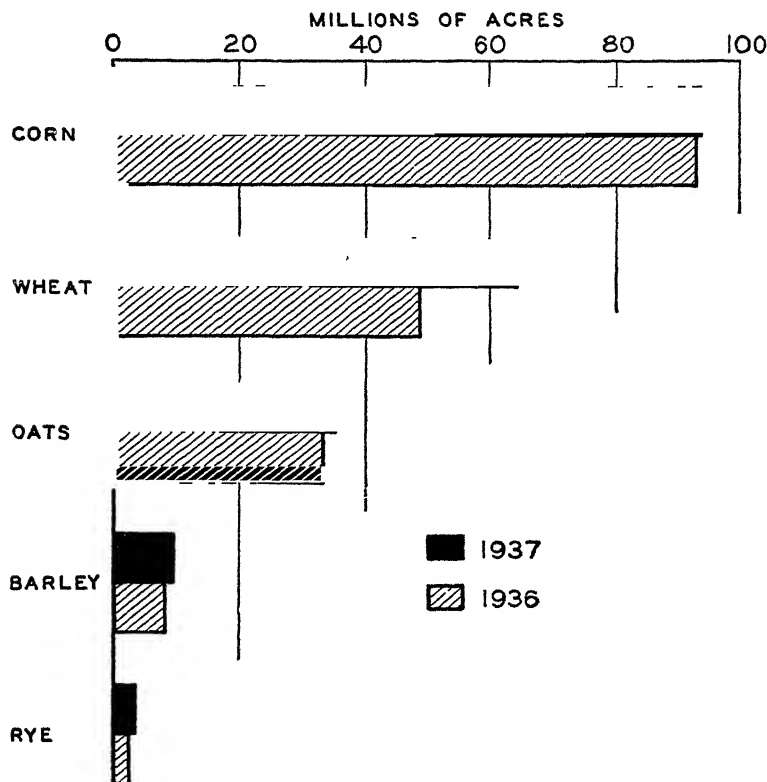
Chart 60.   Acreage Harvested in the United States of Corn, Wheat, Oats, Barley, and Rye, 1936 and 1937.   (Data from *Agricultural Statistics, 1938*, pp. 10, 33, 43, 57, and 67.   Data for 1937 are preliminary.)

operation in 1934. The advantage of this chart is that it shows the information for each of the 93 concerns in a more compact form than could well be done otherwise.

Sometimes we wish to compare two sets of data over a period of several years. This may be done by means of a two-unit bar chart, as shown in Chart 59. Similarly, we may wish to compare several categories for two years; such a comparison is shown in Chart 60. We may also use the
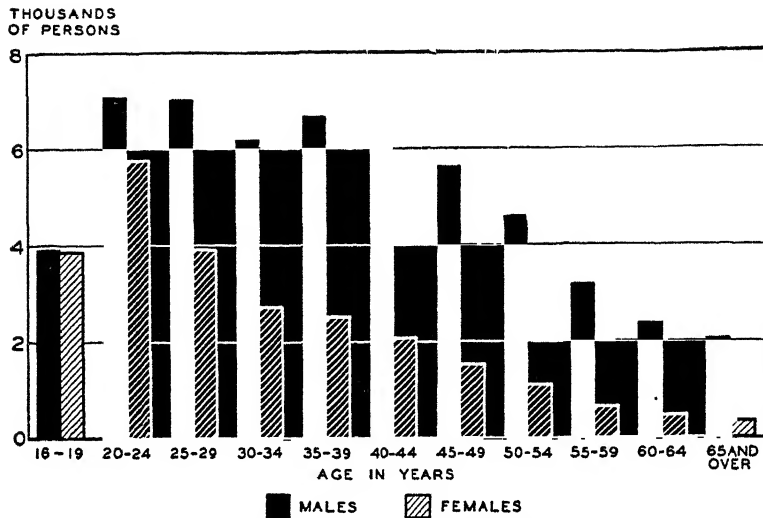
Chart 61. Employable Persons in Philadelphia, by Age Group and Sex, May 1936. (Data from Gladys L Palmer, *Recent Trends in Employment and Unemployment in Philadelphia*, pp. 50, 55, Works Progress Administration, National Research Project in cooperation with Industrial Research Department, University of Pennsylvania, Report No. P-1, December 1937)
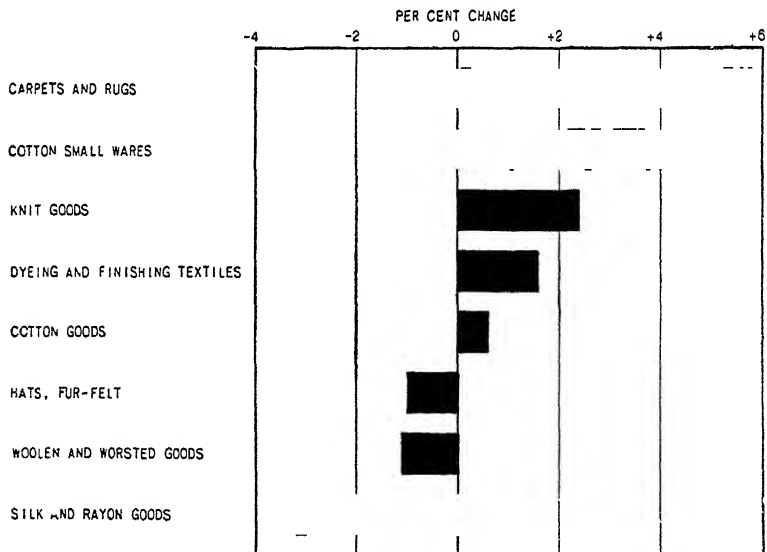
Chart 62. Per Cent Change in Employment in Eight Divisions of the Textile Industry, September–October, 1938. (Based on data from *Monthly Labor Review*, January 1939, p. 237.)

two-unit bar chart to compare several categories each of which is subdi-
vided into two parts, as in Chart 61.

A two-direction bar chart, such as Chart 62, may be used to show in-
creases and decreases. Such a chart is even more effective if increases
can be shown in black and decreases in red. Increases and decreases in
a series of data for a number of years may be shown by means of vertical
bars above and below a horizontal zero line.

Each symbol represents 250,000 tractors   PICTORIAL STATISTICS, INC.
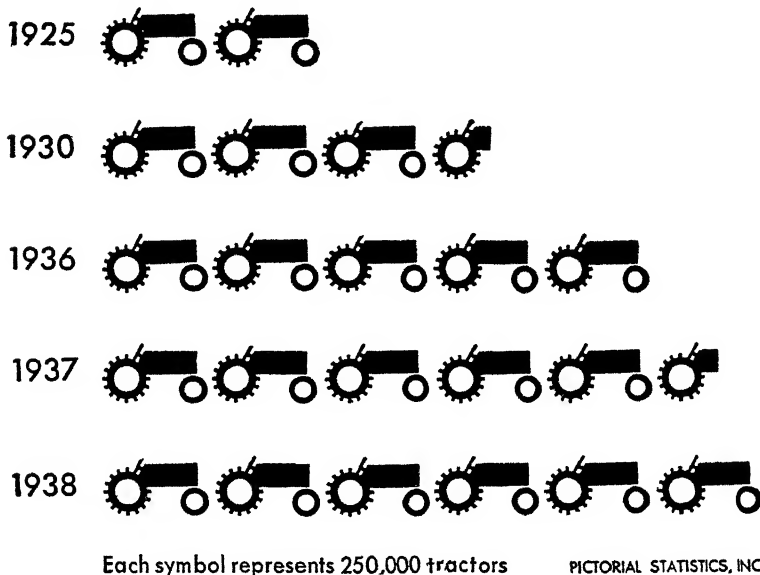
Chart 63.   Tractors on Farms in the United States, 1925, 1930, and 1936–1938, as
Shown by a Pictograph.   (Data from same sources as Chart 54.)

Pictorial Devices

In section D of Chart 54 the number of tractors on farms at each of cer-
tain years was represented by means of pictures of tractors of varying size.
While this sort of chart does not convey a satisfactory comparison to a
reader, it does attract attention.   The pictorial value may be retained by
using a number of small pictures, all of the same size, and arranging them
so as to form a bar chart.   Such a graph is often referred to as a *pictograph*.
Chart 63 shows a comparison of tractors on farms by means of this device.
While the diagram is essentially a bar chart, it is more attractive and thus is
more likely to be examined by a reader.   No scale is used, but since the
pictures are all of the same size and since each represents 250,000 tractors,
approximate numerical values may be had from the chart, if they are

wanted. Although a bar chart of a time series generally uses vertical bars, it will be observed that the pictograph shown as Chart 63 has horizontal bars. Pictographs are often arranged in this way because it seems

CAPACITY AND PRODUCTION OF ELECTRIC POWER PLANTS

1909

1919

1929

1934
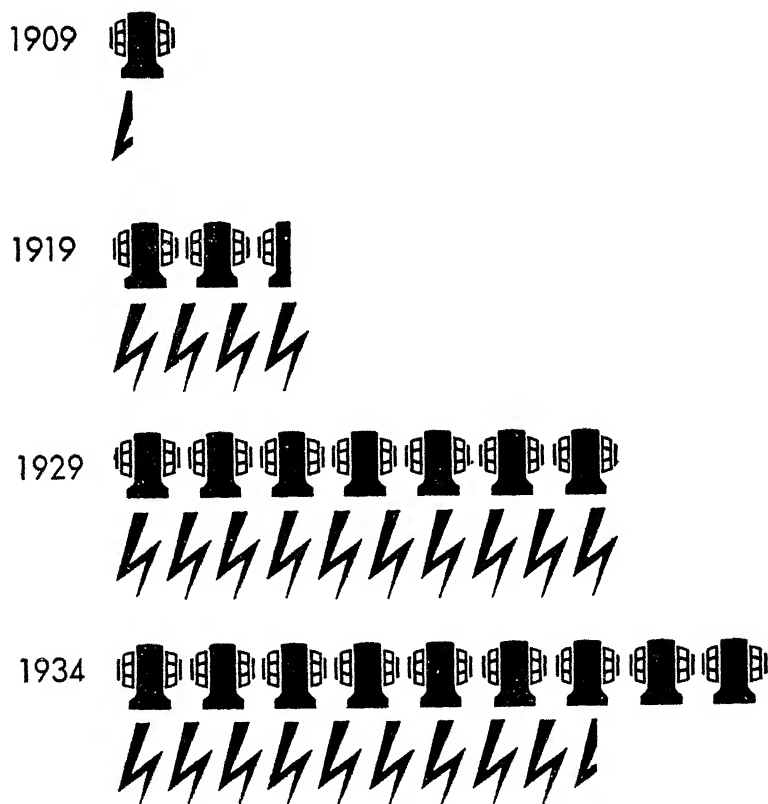
Each dynamo represents 5 million KW capacity

Each bolt represents 10 billion KWH production

Chart 64. A Pictograph from Rudolph Modley, *How to Use Pictorial Statistics* p. 35, Harper and Brothers, New York. 1937.

more suitable to have tractors, people, houses (or whatever is being pictured) standing side by side rather than on top of one another.

Chart 64, another example of a pictograph, shows a comparison of the capacity and production of electric power plants. Chart 65 represents a slightly different application of the idea, in that bars are actually used but the pictures are shown in white against the black background of the bars. It should be apparent that, in making a pictograph, the picture is so chosen
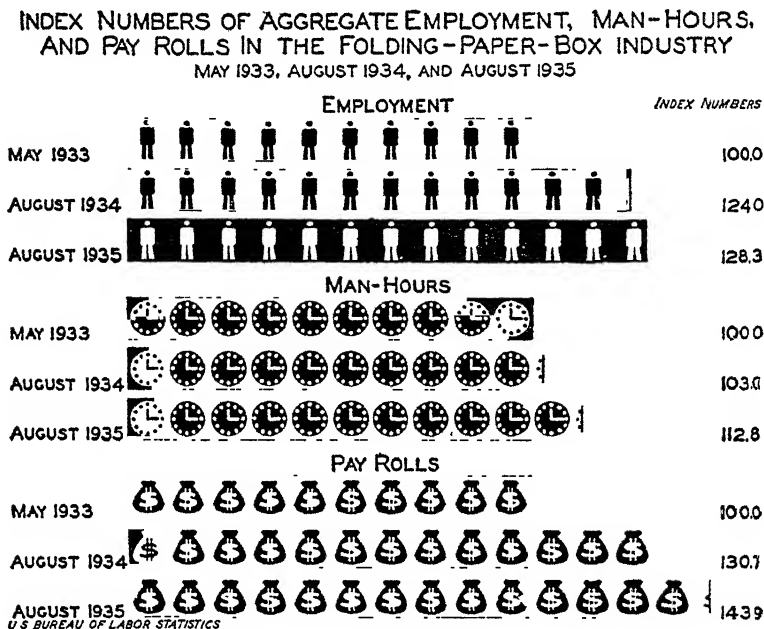
Chart 65. A Modified Pictograph. (From United States Bureau of Labor Statistics, *Wages, Hours, and Working Conditions in the Folding-Paper-Box Industry, 1933, 1934, and 1935*, Bulletin No. 620, p 51.)

as to suggest the nature of the data being shown. Certain basic rules for the use of pictorial devices are shown in Chart 66.

Component Part Charts

The parts of a total may be shown by means of a bar as in Chart 67 or by a pie diagram as in Chart 68. The bar chart involves a one-dimensional comparison of the lengths of the sections of the bar; whereas the pie diagram involves a two-dimensional comparison of the pie sections, or a one-dimensional comparison of the arcs of the pie sections, or a comparison of the central angles. Accuracy of judgment is about the same whether

based on a bar chart or a pie diagram,[2] with the exception that 25 per cent (shown by a right angle) and 50 per cent (shown by a diameter) sections are more accurately gauged from a pie diagram. The pictorial value of

BASIC RULES
FOR PICTORIAL STATISTICS

I SYMBOLS SHOULD BE SELF-EXPLANATORY

II LARGER QUANTITIES ARE SHOWN BY A LARGER QUANTITY OF SYMBOLS

Each ship represents x million tons

NOT BY LARGER SYMBOLS

x million tons   4x million tons
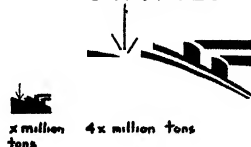
III CHARTS COMPARE APPROXIMATE QUANTITIES

Each ship represents 5 million tons
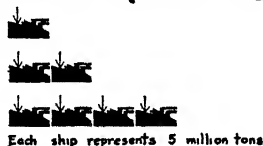
NOT MINUTE DETAILS

4,873,285 tons
11,075,357 tons
20,468,953 tons

IV ONLY COMPARISONS SHOULD BE CHARTED

1870
1900
1930

Each ship represents x million tons
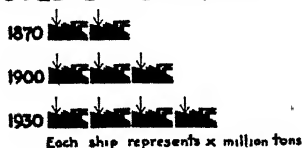
NOT ISOLATED STATEMENTS

1930

Each ship represents x million tons

PICTORIAL STATISTICS, INC.

Chart 66. Basic Rules for Drawing Pictographs as Suggested by Rudolph Modley. (From Rudolph Modley, *How to Use Pictorial Statistics*, p. 15, Harper and Brothers, New York, 1937.)

[2] See "Bar Charts Versus Circle Diagrams," by Frederick E. Croxton and Roy E. Stryker. *Journal of the American Statistical Association*, December 1927, pp. 473–482.

the pie diagram is perhaps greater than that of the bar chart, and it is increased when the pie diagram is designed to suggest a silver dollar.   Chart 69 shows an interesting use made of the pie diagram, which in this case represents 50 cents since that is the fare charged on certain tunnels and bridges operated by the Port of New York Authority. A single component-part bar is occasionally drawn without a scale and is sometimes horizontal.   One advantage of the vertical bar over either the horizontal bar or the pie diagram is that the sections are easier to label (see Chart 67).

Several suppliers of graph paper offer sheets showing a circle with the circumference graduated from 0 to 100, thus enabling us to construct pie diagrams more readily. If such sheets are not available or if varying sizes of circles are desired, pie diagrams may be made by the use of compasses and a protractor.   Since the conventional protractor divides a circle into 360 parts or degrees, the percentages which are to be shown should be multiplied by 3.6.   Dividing a circle into percentages is facilitated by the use of a protractor[3] calibrated to divide a circle into 100 parts, as shown in Chart 70; such a scale may be engraved or otherwise marked on the back of an ordinary protractor (page 137).

Chart 67.   Proportion of the Population of the United States in Each Specified Age Group, 1930. (Data from *Fifteenth Census of the United States, 1930,* Population Volume II, p 576.)

Chart 71 shows how bar charts may be used to compare several sets of component parts and also how the same comparisons may be made by means of pie diagrams.   It seems clear that comparisons between the years are made more easily from the bars than from the circles.   The guide lines running from section to section assist in making comparisons from the bar chart: when the lines are parallel there has been no change; when they diverge, there has been an increase; when they converge, a decrease has occurred.

The comparison of component parts in Chart 71 is on a relative basis; the proportion of each age group in the population is shown.   When we indicate how many of each age group were enumerated, we have diagrams

---

[3] See "A Percentage Protractor," by Frederick E. Croxton, *Journal of the American Statistical Association*, March 1922, pp. 108–109.
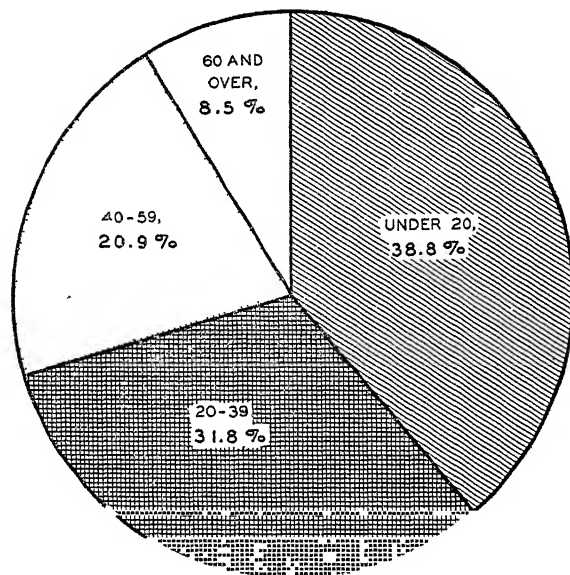
Chart 68.   Proportion of the Population of the United States in Each Specified Age Group, 1930.   (Data from same source as Chart 67.)

such as are shown in Chart 72.   The bars and circles vary in size because the total population has increased.   In this instance the bar chart is clearly preferable to the pie diagram.   When data such as those shown in Charts 71 and 72 cover a number of years, it is generally preferable to make use of curves as was done in Charts 30 and 31.   While the bar charts of Charts
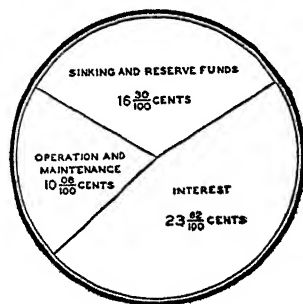
Chart 69.   A Pie Diagram Used by the Port of New York Authority.   This was a 4-page folder about 3½ inches in diameter.   Pictured are pages 1 and 3.   Page 1 was printed in silver and black.   Page 2 showed Port Authority name and address and the statement "The 50¢ toll—where it went, bridges and tunnel, year 1935."   Page 4 gave a brief statement of the expenditure for each of the seven items going to make up "operation and maintenance."

71 and 72 present chronological data, we may also compare component parts for different places or categories. For example, we might compare the proportions of males and females in the urban population with the proportions of males and females in the rural population. One bar, subdivided for males and females, would represent the urban population; the other bar, similarly divided for the sexes, would represent the rural population.

Statistical Maps

Statistical maps are graphic devices which show quantitative information on a geographical basis. We shall consider hatched or shaded maps, dot maps, and pin maps.

Hatched maps. Hatched or shaded maps undertake to show for each geographical area under consideration the magnitude of the phenomenon which is being studied. The variations in magnitude are represented graphically by progressive differences in hatching or shading. In Chart 73 the various hatchings indicate the crop conditions in the drought area of the United States during the period 1930-1936. The counties having relatively poorest crops are

Chart 70. Percentage Protractor.

shown in solid black, and the hatching becomes progressively lighter so that the lightest indicates the counties which had relatively the best crops. Since the drought area did not follow state lines, the parts of six states which were not considered as in the drought area are shown in white. The outstanding characteristic of maps such as this is that a progressive change in the hatching or shading indicates an increase (or occasionally a decrease) in the phenomenon being measured.

Chart 74 shows a shaded map. On this map the states limiting truck drivers to 7 or 8 hours at the wheel are indicated by white; progressively darker shaded areas show the states which permit longer driving periods; solid black indicates no limit.

Sometimes statistical maps are made in colors. However, the principle of progressive shading cannot be developed satisfactorily by using different colors. It is possible, of course, to use progressive shades of a single color and thus sometimes to produce a more attractive map than could be done by using black and white.

Dot maps. The preceding statistical maps were used to show averages

or ratios—average yield per acre and hours per day. When, however, the object is to show the geographical distribution of occurrences, the dot map should be used. Chart 75 shows one of the simplest of dot maps. Each dot represents a service station, and the concentration of them in various parts of the country is clearly shown. In order to avoid heavy concentrations of dots at large centers of population, maps such as this are sometimes made on a county basis so that all the cases in a county are
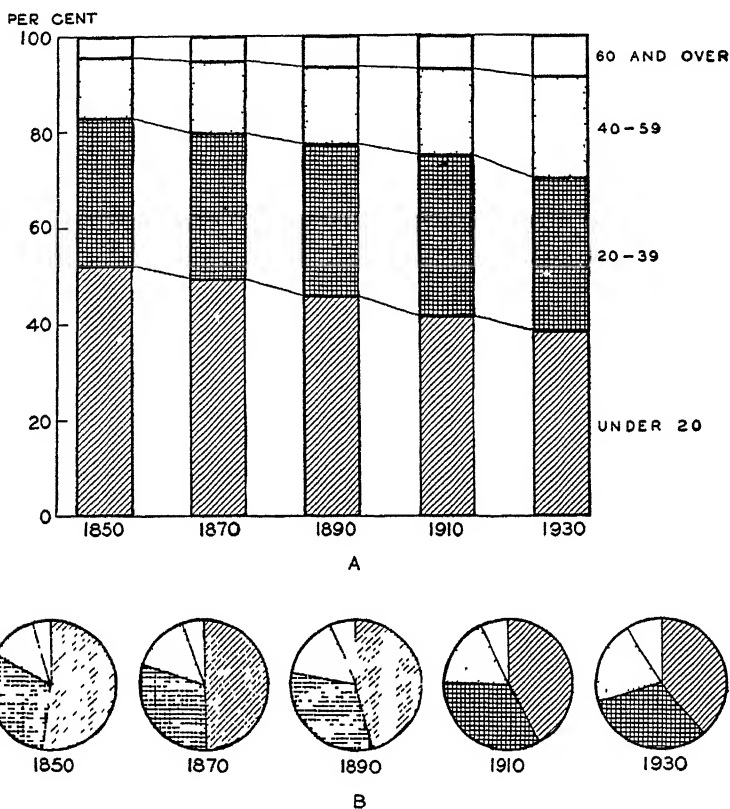
Chart 71. Proportion of the Population of the United States in Each Specified Age Group, 1850, 1870, 1890, 1910, and 1930. If the pie diagrams were shown separately, a legend would be necessary to identify the various hatchings. (Data from same source as Chart 67.)

more or less evenly distributed throughout the county. On the other hand, when the dots are located at the exact place of occurrence, the heavy concentrations of dots, which often become black blotches, indicate clearly the concentrations of occurrences in those areas. Chart 76, another dot map, shows the increase in the number of farms in the United States from

1930 to 1935.   Here the concentrations are clearly evident and show up in several places as areas of solid black.

In drawing a dot map, the number of units represented by one dot may be large, so that the number of dots in a region is small enough to be
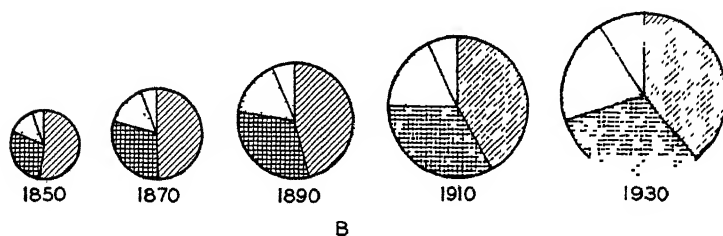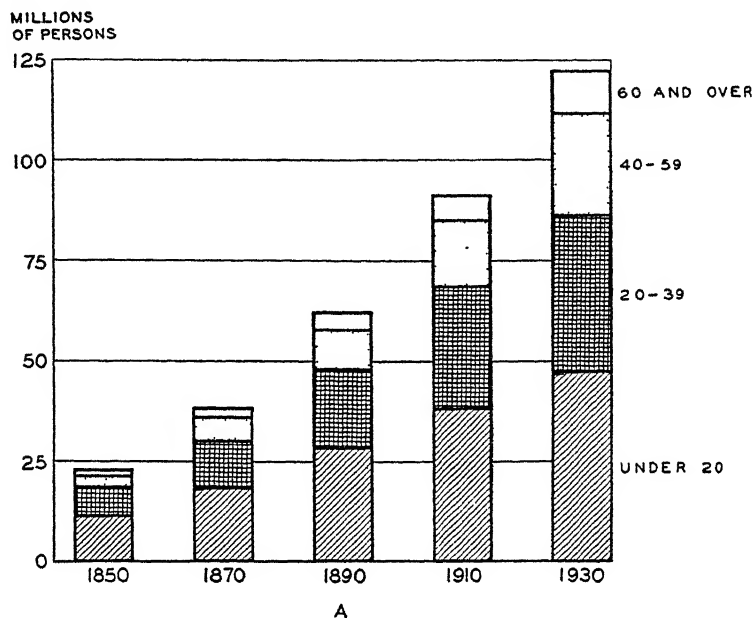
Chart 72.   Population of the United States in Each Specified Age Group, 1850, 1870, 1890, 1910, and 1930.   If the pie diagrams were shown separately, a legend would be necessary to identify the various hatchings.   (Data from same source as Chart 67.)

counted, or the number of units represented by one dot may be small, so that the numerous dots give the effect of a gradual change in intensity of shading from light to dark.   Which technique to use depends on the purpose of the chart.

A different sort of dot map is shown in Chart 77, which uses dots of

Source United States Department of Agriculture   AF-2292 WPA
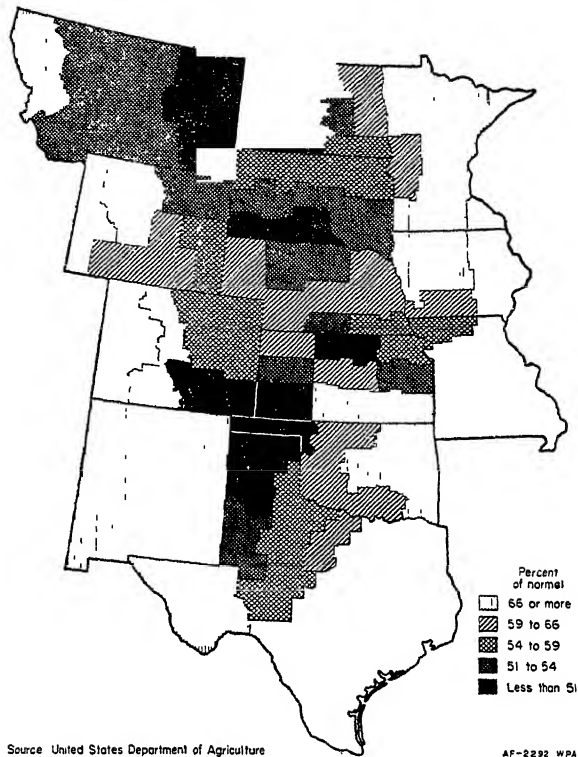
Chart 73. Crop Conditions in the Drought Area. Average crops for 1930–1936 expressed as per cent of normal The white areas were not considered as part of the drought area. (From Works Progress Administration, Division of Social Research, *Areas of Intense Drought Distress, 1930–1936,* p. 15.)
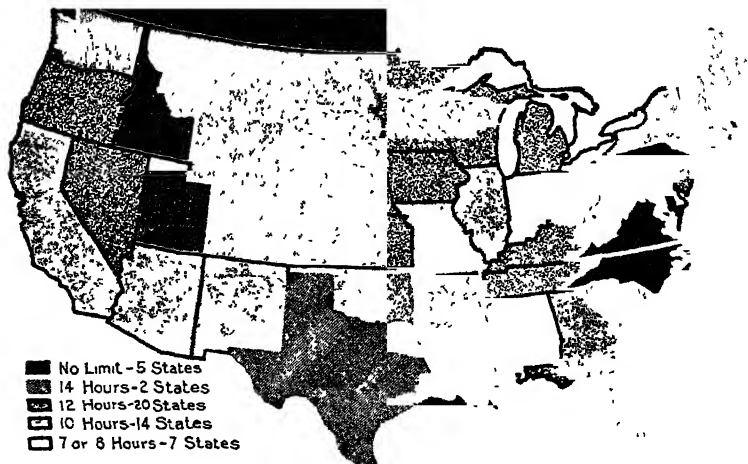
Chart 74. Legal Limit on Number of Driving Hours of Common-Carrier Truck Drivers, by States, 1937. (National Safety Council, *How Long on the Highway,* p. 27.)
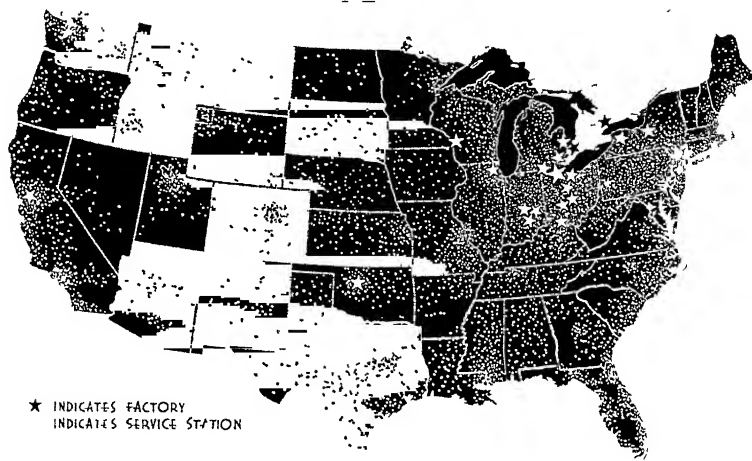
Chart 75. Auto-Lite Factories and Service Stations in the United States, 1936. (Reproduced from an advertisement of the Electric Auto-Lite Company.)

varying size.   In this study 4,030 truck drivers were stopped at various places and were asked how long they had been driving and certain other correlative questions.   The areas of the circles indicate the relative number of drivers questioned at each point.   While the varying circle sizes indicate clearly that more drivers were quizzed at certain places than at others, it is not easy to make accurate comparisons from these dots.   We cannot
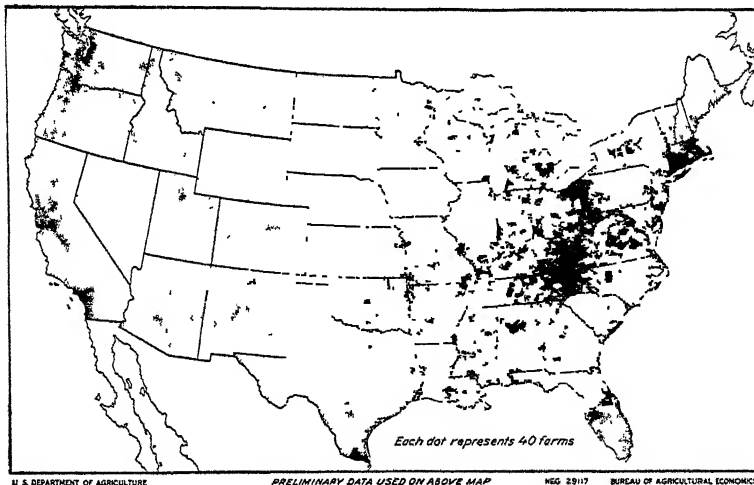
Chart 76.   Increase in Number of Farms in the United States, 1930-1935.   (From the United States Department of Agriculture, Bureau of Agricultural Economics.)
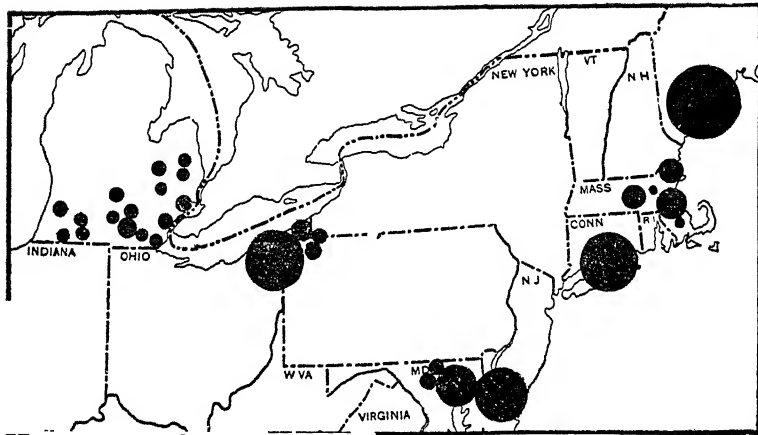
Chart 77. Number of Drivers Interviewed and Location of Interview in a Study of Driving Practices of Truckers, 1936. (Reproduced from National Safety Council, *How Long on the Highway*, p 19 Note that five of the states are not identified.)
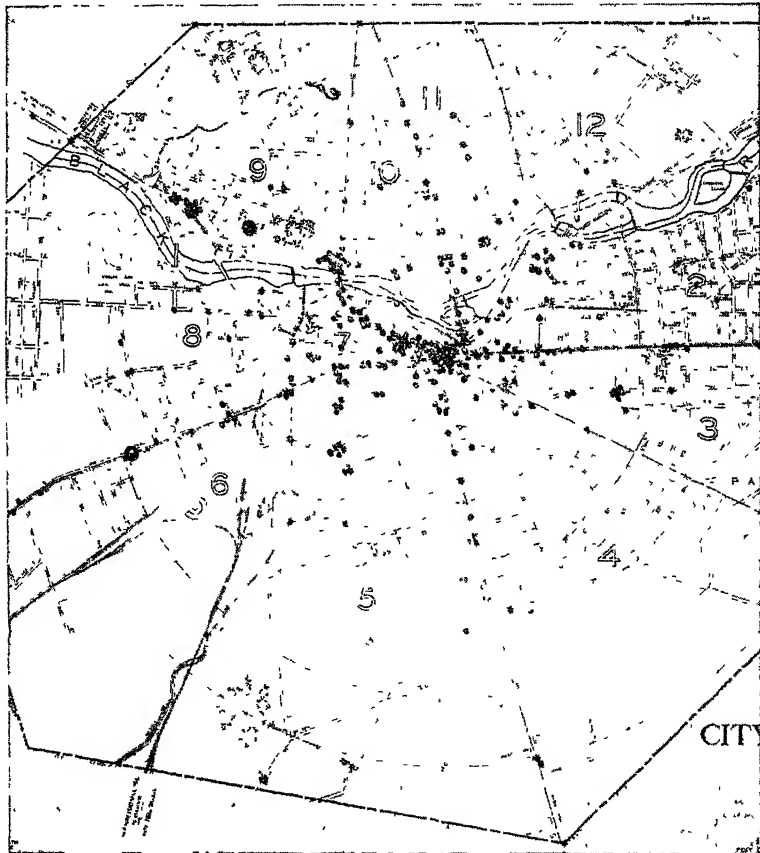
Chart 78. A Portion of an Automobile Accident Map of the City of Watertown New York. The full effect of this map is not apparent in a black and white photograph. The pins used were as follows: black, fatal accident; yellow, pedestrian injured; light blue, automobile and bicycle; red, collision of two or more cars; crystal, light pole broken by automobile. (From National Safety Council. Chicago, Illinois.)
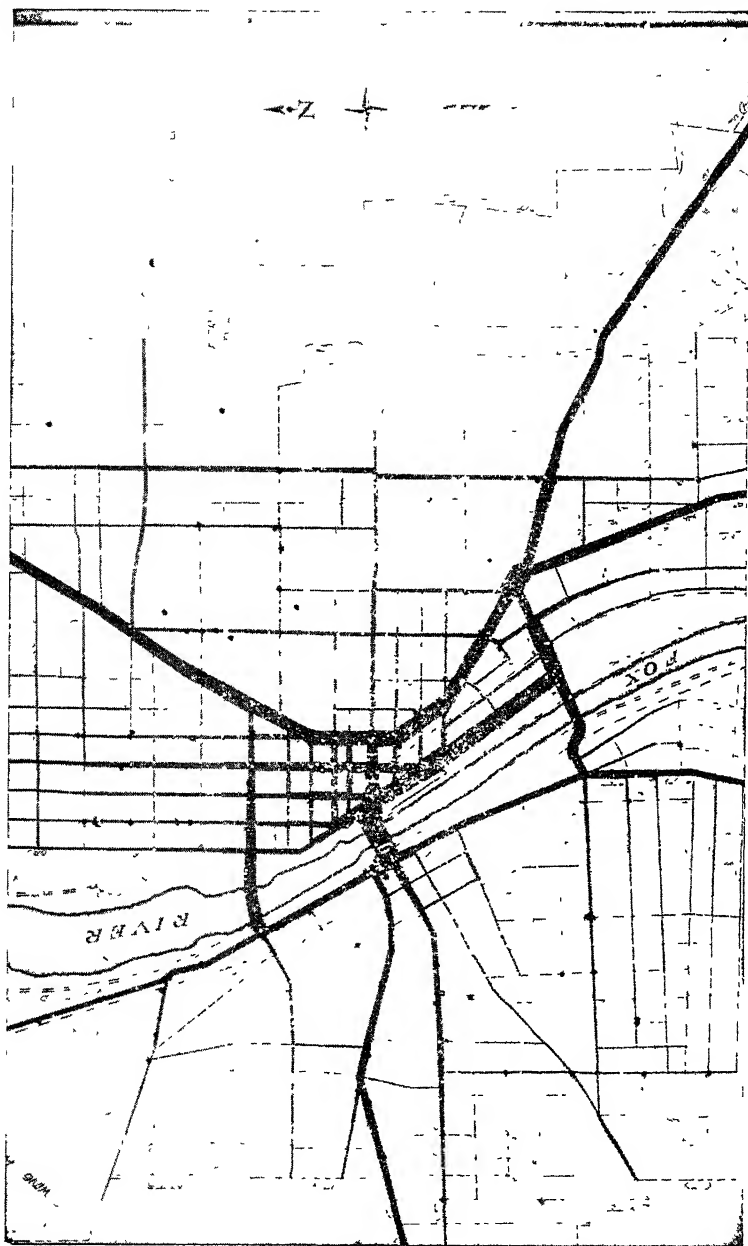
Chart 79. A Portion of a Traffic Flow and Automobile Accident Map of Elgin, Illinois, 1935. The width of the black lines indicates the number of vehicles during the peak hour period. Classes range from 50 to 900 cars per hour. A large nickel pin denotes a fatal accident, a red pin (black in the photograph) indicates a personal injury, a white pin means property damage of less than $10, and a yellow pin designates property damage of $10 or more. White and yellow pins both appear white in the photograph. (Photograph of map furnished by the National Safety Council, Chicago, Illinois.)

143

compare diameters directly.   We must remember that, if one circle has a diameter twice as great as another, then the first circle has an area four times that of the second.

Pin maps.   Pin maps may be thought of as a particularly flexible sort of dot map.   They consist of maps mounted on backing of cork, cardboard, wallboard, corrugated cardboard, etc., on which information is recorded by means of pins having (usually) glass heads of different size, color, and shape.   The available pins range in size from those having heads about $\frac{1}{16}$ inch to about $\frac{3}{4}$ inch in diameter.   A large number of colors is available
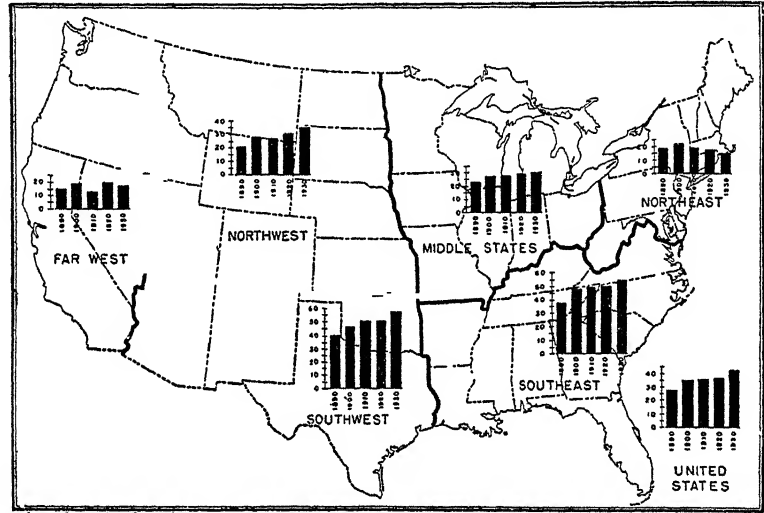
Chart 80.   Percentage of Farms Operated by Tenants by Regions of the United States at Each Census, 1890-1930.   (Reproduced from National Resources Committee, *Problems of a Changing Population*, p. 61.)

as well as round-, square-, and triangular-head pins.   Pin maps may be readily altered as the facts change.   Because of this flexibility and the wide variety of pins available, the pin map is a very useful means of presenting geographical data.   An extensive pin map scheme, involving one or more maps mounted on cork and hundreds or thousands of pins, is expensive but may often prove very useful.

Charts 78 and 79 show two examples of pin maps used to record automobile accidents.   By using one or more such maps, it is possible not only to observe the frequency with which accidents occur at various places, but also the nature of each accident (automobile hitting pedestrian, automobile hitting automobile, automobile hitting fixed object, etc.) and the result of the accident (property damage, occupant injured, occupant killed, pedestrian injured, pedestrian killed, etc.).

One difficulty with the statistical map is that the importance of different regions is not to be judged by their areas. For instance, a hatched map showing income per family in different states would be somewhat misleading because there are many more families in some of the states occupying a very small area than there are in other states occupying a very large area. An interesting device for overcoming this difficulty is to draw the map in such a way that the area of each state is in proportion to its population. Chart 230, page 741, shows such a map.

Occasionally a map and some other type of chart are used in combination. Chart 80 shows such a usage. The data to be presented consisted of the percentage of farms operated by tenants at each of five censuses and for each of six parts of the United States as well as for the country as a whole. With the seven bar charts placed on the map, the reader may visualize exactly what territory is referred to in each instance.

Selected References

H. Arkin and R. R Colton. *Graphs How to Make and Use Them*, Chapters VI, VII, IX; Harper and Brothers, New York, 1936.

W. C. Brinton: *Graphic Methods for Presenting Facts*, Chapters I–IV, XI–XV; McGraw-Hill Book Co., New York, 1914.

W. C. Brinton. *Graphic Presentation*, Chapters 10–29, Brinton Associates, 608 West 45th St., New York, 1939.

A. C. Haskell: *Graphic Charts in Business* (Second Edition), Chapters XII–XV, XXI–XXVI; Codex Book Co., Norwood, Mass., 1928.

K. G. Karsten: *Charts and Graphs*, Chapters I–XIV, XXIII–XXV, L–LVI; Prentice-Hall, Inc , New York, 1923.

R. Modley. *How to Use Pictorial Statistics;* Harper and Brothers, New York, 1937. A guide to the construction and use of pictographs.

J R. Riggleman· *Graphic Methods for Presenting Business Statistics* (Second Edition), Chapters III, VIII; McGraw-Hill Book Co., New York, 1936.

CHAPTER VII

RATIOS AND PERCENTAGES

It was pointed out in the chapter dealing with statistical tables that derived figures are useful to assist in summarizing and comparing data. In that chapter specific mention was made of ratios, percentages, and averages. This chapter will discuss ratios and percentages. Averages and related measures will be examined in later chapters.

To express the ratio which 753 bears to 251, we divide 753 by 251, which gives 3, and we say that 753 is to 251 as 3 is to 1, or more briefly 753 : 251 : : 3 : 1. We have thus indicated the relationship which the first of these two numbers bears to the second as a *ratio to one*. If it suited our purpose better, we could express the relationship as a ratio to any other number. For example, we could use a ratio to ten, saying 753 : 251 : : 30 : 10; we could use a ratio to one hundred and write 753 : 251 : : 300 : 100. This last ratio, per hundred, is generally referred to as a percentage, and we note that 753 is 300 per cent (from *per centum*) of 251. It will thus be seen that percentages, which are used so frequently, are merely special cases of the more general concept of ratios. If, instead of using a ratio per hundred, we find occasion for a ratio per thousand we may refer to our figures as "per mille."

Ratios are computed in order to expedite comparisons. Not only are large numbers reduced as in Table 3, but much is gained by comparing a series of figures with a rounded base of 100 (which can be carried in one's mind) rather than by attempting to compare each individual economic class of exports with total exports and each economic class of imports with total imports. Relative change may be visualized more concretely when shown by percentages as in Table 20, or when shown by one of the methods in Table 21.

Calculation

When one or more numbers are being compared to another number, the figure to which comparisons are made is known as the *base*. A ratio is found by dividing the figure, which is being compared to the base, by the

TABLE 20

PRODUCTION AND YIELD PER ACRE OF SELECTED GRAINS IN THE UNITED STATES
1936 AND 1937

Grain	Production (thousands of bushels)		Per cent increase in production	Yield per acre (bushels)		Per cent increase* in yield per acre
	1936	1937		1936	1937	
Corn .. .	1,507,089	2,644,995	75 5	16 2	28.2	74.1
Oats .	785,506	1,146,258	45 9	23 5	32 7	39.1
Wheat	626,766	873,993	39 4	12 8	13.6	6 2
Barley	147,475	219,635	48 9	17 6	22.1	25.6
Grain sorghums	55,079	97,097	76.3	8.0	13.2	65 0
Rice .	49,002	53,004	8 2	50 6	48.5	−4.2
Rye	25,319	49,449	95 3	9 1	12.9	41 8
Buckwheat	6,285	6,777	7 8	16.8	15.9	−5.4

* A minus sign denotes a decrease
Source *Crops and Markets*, Vol 14, No 12, December 1937, pp 259–261

base. (The use of calculating machines is discussed in Appendix C.)
The figure is then expressed in terms of or in relation to the base, and ratios
of all sorts are therefore sometimes referred to as *relative numbers* or *rela-
tives*.

TABLE 21

PRODUCTION OF POTATOES IN THE UNITED STATES, 1926–1937

Year	Production (thousands of bushels)	Per cent of 1926	Per cent increase over 1926	Per cent of preceding year	Per cent increase* over preceding year
1926	321,607	100 0	..	.	..
1927	369,644	114 9	14 9	114.9	14.9
1928	427,249	132.8	32 8	115.6	15 6
1929	332,204	103 3	3.3	77.8	−22.2
1930	340,572	105.9	5 9	102.5	2.5
1931	384,125	119 4	19.4	112.8	12.8
1932	376,425	117 0	17.0	98.0	− 2.0
1933	342,306	ʹ106.4	6 4	90.9	− 9 1
1934	406,105	126.3	26.3	118.6	18.6
1935	386,380	120 1	20 1	95.1	− 4.9
1936	331,918	103 2	3.2	85.9	−14 1
1937	391,159	121.6	21 6	117.8	17.8

* A minus sign denotes a decrease
Source *Crops and Markets*, Vol 14, No. 12, December 1937, p 261.

The amount of money in circulation in the United States on June 30,
1914, was $3,459,434,174.   On June 30, 1938, the circulating medium to-
taled $6,461,058,390.   To state the 1938 circulation in terms of the 1914

circulation (the base), we divide $6,461,058,390 by $3,459,434,174 and obtain 1 868.   This figure means that the circulation in 1938 was 1.868 times as great in that year as in 1914.   In many instances ratios are most useful when given as percentages.   To change 1.868, the ratio to one, to a ratio per hundred the decimal point is moved two places to the right; the resulting figure, 186.8 indicates that money in circulation in 1938 amounted to 186.8 per cent of that in 1914.

It should be noticed that there are two ways in which we can express the percentage figure just arrived at.   Instead of saying that the 1938 circulation was 186.8 per cent *of* 1914 circulation, we may state that circulation in 1938 was 86.8 per cent *greater* than in 1914.   In the first instance we compared the totals for the two years; in the second we compared the change which took place with the 1914 total.[1]

Effect of Changing Base

Naturally a different set of percentages would be arrived at if we compared the 1914 circulation figures to the 1938 figures.   We are now using 1938 as the base, and the 1914 figure is divided by that for 1938.   Performing this operation indicates that circulation in 1914 was 53.5 per cent of that in 1938, or that circulation in 1914 was 46.5 per cent less than that in 1938.   Observe that, while the 1938 figure was 86.8 per cent greater than the 1914 figure (1914 was the base), the 1914 figure was 46.5 per cent less than the 1938 figure (1938 was the base).   This difference is, of course, due to the fact that the basis of comparison was first in reference to 1914, then to 1938.   If a number is increased 100 per cent, the second number need be decreased but 50 per cent to arrive at the original figure.   Conversely, if a given number is decreased 50 per cent, the second number must be increased 100 per cent to reproduce the given number.

The failure to realize the effect of this change of base may lead to the drawing of false conclusions.   Some years ago a firm decreased the wages of its employees 15 per cent; later it increased the reduced wages 5 per cent; then it raised these increased figures 5 per cent; and finally it increased these second figures another 5 per cent.   Afterwards it announced that the three 5 per cent increases put wages back where they were before the 15 per cent reduction.   Calculation will show that the new wages were really 98.4 per cent of the original wages before reduction.   If the com-

---

[1] Suppose we are comparing two percentages, as 4.0 per cent and 9.0 per cent.   We may speak in absolute terms and say that 9.0 per cent is 5.0 per cent more than 4.0 per cent.   We may speak in relative terms and say that 9.0 per cent is 125 per cent greater than 4.0 per cent, or that 9.0 per cent is 225 per cent of 4.0 per cent.   When comparing percentages, it is advisable to make quite clear whether we are speaking in absolute or relative terms.

pany had given a single 15 per cent increase of the reduced wages, the new wages would have been but 97.75 per cent of the original wages.

Table 22 shows for selected percentages of increase the per cent which the new number must be decreased to reproduce the original number. It should be borne in mind that a per cent of increase figure may be indefinitely large; however, a per cent of decrease figure of 100 indicates a decline to zero, while a per cent of decrease of over 100 indicates a fall to a negative quantity.

Recording Percentages

Generally percentages are recorded to one decimal place. If the percentages are based upon large figures and particularly if one part of a total is quite small (see Tables 7 and 24), it may be desirable to use more than one decimal. Occasionally only whole percentages are shown and enable relationships to be grasped readily. Whole percentages will not suffice, however, when the relative variations are extremely small.

TABLE 22

ILLUSTRATIONS OF EFFECT OF SHIFTING BASE IN CALCULATING PERCENTAGES

Given number	Per cent of increase	New number	Per cent new number must be decreased to yield given number
10	500.00	60 00	83 33
10	200 00	30 00	66 67
10	100 00	20 00	50 00
10	50.00	15.00	33.33
10	33 33	13 33	25.00
10	25 00	12 50	20 00
10	10 00	11.00	9 00
10	5 00	10.50	4 76
10	1 00	10 10	.99

Percentages should not be calculated if the absolute numbers are small, especially if the base is appreciably less than 100. A serious difficulty arising out of the use of percentages based on small absolute numbers is discussed on pages 160–161.

When percentages are to be recorded with one decimal, they are correct to the nearest tenth of one per cent. The following examples will indicate the procedure in rounding percentages (and also in rounding other calculations involving remainders):

(1) $371.16 ÷ $679.28 = .5464, or 54.64 per cent. The second decimal is less than 5 and therefore this percentage, to the nearest tenth of one per cent, is 54.6.

(2) 2,319 pounds ÷ 7,532 pounds = .3079, or 30.79 per cent. In this instance the second decimal is more than 5 and the percentage should be recorded as 30.8.

(3) 280,511 feet ÷ 11,000,000 feet = .025501, or 2.5501 per cent. Here the second decimal is 5, but there is a remainder which results in the 1 in the fourth decimal place. Recorded to the nearest tenth of one per cent this figure is 2.6.

(4) 1,341 barrels ÷ 6,000 barrels = 2235, or 22.35 per cent. Here the nearest tenth is either 22.3 or 22.4. It does not greatly matter whether occasional results such as this are raised in the first decimal place or whether the second decimal is dropped. However, it is better to follow some consistent scheme. Particularly when many ratios are being calculated which are eventually to be added, it is well to employ a method which will cause half of the ratios with a second decimal of exactly 5 to be raised and half to be lowered. This practice will avoid the accumulation of errors. Probably the most satisfactory scheme is to raise the first decimal when the first decimal is an odd number (67.35 becomes 67.4) and to drop the second decimal when the first decimal is an even number (67.65 becomes 67 6).

Reference to the percentage data shown in Table 4 will reveal that the nine percentages add to 100.1 rather than to 100.0. This is the consequence of rounding all percentages to one decimal place, which fairly often results in totals of 99.9 or 100.1 and occasionally shows 99.8 or 100.2. Some statisticians adjust one of the percentages in order to produce the correct total, but it seems preferable to let each percentage stand correctly rounded, as in Table 4. It is interesting to note that, if the individual percentages are carried to one more decimal place than is the total, this apparent discrepancy does not occur.

Types of Comparisons

We have already seen instances in which the parts of a whole are compared to the total in Tables 3, 4, and 7. Here the percentages were obtained by dividing each item in turn by the total. More expeditiously we may take the reciprocal of the total and multiply the reciprocal by each of the component figures. This is a time-saving device adapted particularly to the calculating machine, and is applicable whenever we are dividing a series of numbers by a constant number.

An example in which one part of a total is compared with another part of another total is given in Table 8. In this table each figure for males was divided by the appropriate figure for females, since the sex ratio consists in stating the number of males per 100 females.

Table 21 indicates a number of different comparisons which may be made

in regard to data arranged chronologically.   In column 3 the production
of potatoes for each year is compared with the 1926 production; each figure
is divided by that for 1926.   Column 4 shows the percentage by which
the production for each year exceeded that for 1926; each year's nu-
merical increase or decrease over 1926 is divided by the 1926 production.
In column 5 the production each year is related to that of the preceding
year; each year's figure is divided by that for the preceding year.   Column
6 indicates the per cent of increase or decrease of each year in relation to
the preceding year; the numerical increase (or decrease) of each year over
the preceding year is divided by the production for the preceding year.
In columns 3 and 4, comparisons are made with a fixed base, 1926.   In
columns 5 and 6 the base is constantly shifting, being always the preceding
year.

Another application of percentages is shown in Table 20.   Here the
1936 figure for each crop is the base   The percentage columns headed
"per cent increase" indicate the relative increase or decrease in each crop
from 1936 to 1937.

Some Frequently Used Ratios

The following paragraphs indicate a few interesting applications of ratios
and percentages.   The reader will doubtless become aware of many others
as he reads more or less technical material in magazines, newspapers,
books, and advertisements.

Index numbers.   Most index numbers are presented in the form of per-
centages.[2]   In the construction of an index number of wholesale prices,
for example, the commodities to be included are selected first, and their
prices are then combined with due regard to the varying importance of
the different commodities.   If the index number is a chronological one, as
is usually the case, some year may be designated as the base and prices in
that year are set equal to 100.   The prices for the other years are then ex-
pressed in relation to that base year.   The United States Bureau of Labor
Statistics uses 1926 as the base year for its index numbers of 813 wholesale
prices.   Wholesale prices in 1926 are therefore represented by 100.   The
index number for 1928 was 96.7; for 1929 it was 95.3; it fell to 64.8 in 1932,
rose to 86.3 in 1937, and dropped to 78.6 in 1938.   Prices in each of these
later years are thus expressed in terms of 1926, which is regarded as a repre-
sentative or "normal" year.

Sex ratio.   The relationship of the number of males to the number of
females in the population is given by the sex ratio, which states the number
of males per 100 females.   In 1930 there were 62,137,080 males and

---

[2] See Chapters XX and XXI for a more complete discussion of index numbers.

60,637,966 females in the United States. There were thus 102.5 males per 100 females in the United States, as shown in Table 8. This ratio varied in the different states. It was lowest in Rhode Island, where there were 95.2 males per 100 females, and highest in Nevada, where there were 140.3 males per 100 females. The various nativity groups in the population are listed in Table 8 Negroes showed 97.0 males per 100 females, native whites 101.1 males per 100 females, foreign-born whites 115.1 males per 100 females, Japanese 143.3 males per 100 females, Chinese 394.7 males per 100 females, and Filipinos 1437.7 males per 100 females.

Population density. Instead of merely comparing the total population of two communities, it may often be more meaningful to consider the density of the population. We do this by dividing the total population by the area in square miles, and thus determine the number of persons per square mile. For example, in 1930 the population of Montana was 537,606 and the population of New Hampshire was 465,293. If we relate these figures to the area of each state, we find that New Hampshire had 26.7 persons per square mile, while Montana had but 3.7 persons per square mile. These figures do not, of course, mean that there were 26 or 27 persons on *every* square mile in New Hampshire and 3 or 4 persons on *every* square mile in Montana. They are merely summary figures indicating that, on the average, there were the indicated number of persons per square mile in the state.

Population density may also be used in making chronological comparisons. As our country has grown older, the population density has increased. In 1800 there were 6.1 persons per square mile in the United States; in 1930 there were 41.3 persons per square mile.

Persons per family. With a decline in birth rates there is an accompanying decrease in the size of families. Thus in 1920 there were 24,351,676 families in the United States and a total population of 105,710,620. The average number of persons per family was thus 4.34 in 1920. At the following census there were 29,979,841 families and a total population of 122,775,046. The average family in 1930 was thus composed of 4.10 persons. The term "family" as used here included 75,178 quasi-family groups (institutions, hotels, etc.) in 1930. Quasi-family groups were included but not separately counted in 1920.

Ratios per capita. Many figures are more meaningful or more useful when expressed on a per capita basis. The costs of government in the various states reflect not only the level of expenditure and government services but also the population of the states. For example, the cost of operation and maintenance of the general departments of the State of New York in 1937 amounted to $335,965,861, while in New Jersey it totaled $85,106,172. If these figures are each divided by the population of the

respective state, it appears that the cost was $25.95 per capita in New York and $19.63 per capita in New Jersey.

The consumption of various commodities is frequently stated on a per capita basis. Thus in the period July 1936–June 1937 the "apparent consumption" of oleomargarine (amount withdrawn for consumption) was 3.0 pounds per capita; the estimated consumption of rice was 6.0 pounds per capita, the approximate amount of refined sugar consumed (available for consumption) was 97.4 pounds per capita. The apparent consumption of beef, veal, mutton, lamb, and pork was 126.8 pounds per capita during the calendar year 1936.

The chronological comparison of figures is also frequently facilitated by relating them to the population. On June 30, 1926, the amount of money in circulation was $4,885,266,000. By June 30, 1938, this figure had increased to $6,461,058,390. During this same period, however, the population had been increasing so that the money in circulation had to serve a larger group of people. Expressing the money in circulation in terms of the population, we find that the per capita money in circulation amounted to $41.71 in 1926 and $49.67 in 1938.

Death rates. The crude, gross, or general death rate for a given year is obtained by dividing the number of deaths occurring in a community during that year by the mid-year population of that community, and expressing the result in terms of per thousand. In 1936 there were in the United States 1,479,228 deaths from all causes. The 1930 census, taken as of April 1, 1930, enumerated 122,775,046 persons; and the June 30, 1936, population was estimated to be 128,429,000. The death rate for 1936 was therefore 1,479,228 ÷ 128,429,000 = .0115, or 11.5 per thousand. It will be seen that the accuracy of a death rate depends first upon the degree of completeness of the registrations of deaths, and second upon the accuracy of the mid-year population estimate used as a base. Since population counts are made only once in 10 years, most of the population figures used must be estimates. When the population is estimated for a year falling between two censuses, the estimate is termed an *inter-censal* estimate; when the estimate is for a year after a census, it is termed a *post-censal* estimate. Inter-censal estimates are naturally somewhat more accurate than post-censal estimates. For the years 1931 to 1939 inclusive, death rates must at present be based upon post-censal estimates and are called *preliminary* rates. After the 1940 census results are available, inter-censal estimates may be made for the years 1931–1939, and the death rates may be recomputed upon the basis of these new population estimates. Such rates are called *revised* rates.

When the deaths occurring in a state or city are divided by the population of that community, the resulting crude death rate is subject to certain

corrections. For example, in any given year people may die in a community who are residents elsewhere and also some residents of any large community may die outside of that community. If the non-resident deaths are deducted from those which occurred in the community, the resulting rate is referred to as a *local* rate. If, in addition, the deaths of residents occurring outside of that community are added, the resulting rate is referred to as a *resident* rate. Failure to recognize these important differences may lead to drawing false conclusions. In February 1935 it was announced that the death rate for Queens, borough of New York City, was 6.5 per 1,000, for Bronx 7.8, for Brooklyn 9.3, for Richmond 13.5, and for Manhattan 16.3. The death rate for Queens was lower than for any other such community in the United States, and some persons promptly announced that Queens was "the healthiest place in the country." It was very quickly pointed out in the press, however, that Queens possessed a very low quota of hospitals and that, therefore, some residents of Queens in need of hospital care would seek it in Manhattan or elsewhere. Hospital cases naturally show a very high death rate, and a crude death rate would not reflect the fact that some persons dying in Manhattan and elsewhere were really residents of Queens.

Death rates for particular groups of the population (males and females, various age groups, etc.) and for particular diseases or causes are referred to as *specific* death rates. Because the deaths from any one cause are relatively few, specific rates are usually stated per 100,000 of the population. Thus in 1936 the death rate for diphtheria was 2.4 per 100,000.

An intelligent comparison of the death rates of different communities involves the necessity of adjusting for the fact that the proportions of the sexes may differ, for differences in the age distribution of the population, for variations in the racial and nativity composition of the inhabitants, for differences in occupations, and for other factors. A discussion of these differences and the methods of computing *adjusted* death rates is too specialized a topic to be treated in this text.[3]

Birth rates. Birth rates are usually calculated by dividing the births during a year by the mid-year population for that year. Just as in the case of death rates we may have preliminary rates and revised rates. We may also have gross, local, and resident rates. Stillbirths are not counted as births, although they have been so counted in the past; this fact should

---

[3] See George Chandler Whipple, *Vital Statistics*, Chs. VIII, X, and XII, second edition; John Wiley and Sons, Inc., New York, 1923. The adjustment for age distribution, for example, consists of determining what the death rate would have been if the ages were those of the "standard million" (based on the population of England and Wales in 1901). The death rate for each age group as observed in a community is applied to the number of persons in the corresponding age group of the standard million and the total of these "computed deaths" is related to 1,000,000 to give the death rate adjusted to the standard age distribution.

be remembered in making chronological comparisons. Perhaps it is also worth while calling attention to the fact that the registration of births is not so complete as is the registration of deaths. A death must be registered before a burial permit may be issued and before interment may be made. A newborn infant, however, may be absorbed into the family and the community whether or not his birth is registered.

The calculation of birth rates in relation to the total population is not thoroughly satisfactory since the proportion of "child producers" in the population is not constant either from time to time or from place to place. Refinements in the calculation of birth rates involve the separation of legitimate and illegitimate births, the comparison of legitimate births to the number of married women of child-bearing age, and the comparison of illegitimate births to the total population or to the number of unmarried women of child-bearing age.[4]

Crop yields per acre. Data of the total amount of a crop produced may tell us whether or not there is more of that commodity available in one year than in another. From such figures, however, we cannot know if an increase may have been due to a more abundant yield or to an increase in acreage. In 1933 there were 528,975,000 bushels of wheat harvested from 47,910,000 acres in the United States; in the following year 42,235,000 acres yielded 496,469,000 bushels. Although the total crop was smaller, the yield per acre had risen. In 1933 it was 11.0 bushels per acre, while in 1934 it was 11.8. The yield per acre was 12.2 bushels in 1935, 12.8 bushels in 1936, and 13.6 bushels in 1937. On a geographical basis it is interesting to note that in 1937 the yield varied from 5.6 bushels per acre in South Dakota to 25.6 bushels per acre in Nevada. Our leading wheat states (Kansas and North Dakota, for example) did not show the greatest yield per acre.

Hog-corn ratio. The hog-corn ratio is the result of dividing the average price per 100 pounds which farmers receive for hogs by the average price per bushel which farmers receive for corn. For example, if, as in March 1938, farmers are receiving $8.35 per 100 pounds for hogs and $.513 per bushel for corn, the ratio is $8.35 ÷ $.513 = 16.3. This ratio may be interpreted to mean that 100 pounds of hogs are 16.3 times as valuable as a bushel of corn or, more simply, that 16.3 bushels of corn are equal in value to 100 pounds of hogs. In April 1938 hogs brought $7.77 per 100 pounds and corn yielded the farmer $.527 per bushel. At that time the ratio was 14.7. Over the 26-year period 1910–1935 the hog-corn ratio averaged about 11.1, falling as low as 6.0 in December 1934 and reaching

---

[4] For a more complete discussion of birth rates, see Whipple, *ibid.*, pp. 246–251. The author also discusses marriage rates, divorce rates, morbidity rates, fatality ratios, and so forth.

18.7 in June 1926.   When the ratio is low, it is more profitable for farmers to sell their corn outright than to feed the corn to hogs being fattened for market.   When the ratio is high, it becomes more profitable for the farmer to feed corn to his hogs than to sell the corn outright.   Since corn is the principal element of cost in producing hogs for market, the ratio is used as an indicator of the desirability of future expansion or contraction of hog production.   There is thus a relationship between the hog-corn ratio and the hog production cycle.   When the ratio is high, an increase in hog production tends to follow.   Such an increase is frequently followed by a decline in hog prices in relation to corn prices, and there then follows a tendency to restrict hog production.   Curves showing hog-corn ratios are shown in Charts 49 and 256.

TABLE 23

INDIVIDUAL BATTING AVERAGES OF 22 OUTSTANDING AMERICAN LEAGUE PLAYERS, 1937

Player and club	Games	Times at bat	Hits	Batting average*
Gehringer, Charles L , Detroit	144	564	209	371
Gehrig, Henry L., New York	157	569	200	.351
DiMaggio, Jos. P., Jr , New York	151	621	215	.346
Bonura, Henry J , Chicago .	116	447	154	.345
Travis, Cecil H., Washington	135	526	181	344
Bell, Roy C , St Louis	156	642	218	.340
Greenberg, Henry, Detroit	154	594	200	.337
Walker, Gerald H , Detroit ..	151	635	213	.335
Dickey, William N , New York .	140	530	176	.332
Fox, Ervin, Detroit .	148	628	208	.331
Stone, John T , Washington . .	139	542	179	330
West, Samuel F., St Louis	122	457	150	.328
Selkirk, George A , New York .	78	256	84	328
Vosmik, Joseph F , St. Louis .	144	594	193	.325
Radcliff, Raymond A , Chicago .	144	584	190	.325
Solters, Julius J., Cleveland .	152	589	190	.323
Moses, Wallace, Jr., Philadelphia	154	649	208	.320
Henrich, Thomas D , New York .	67	206	66	.320
Appling, Lucius B , Chicago	154	574	182	317
Allen, Ethan N., St. Louis. . .	103	320	101	.316
Pytlak, Frank A , Cleveland ..	125	397	125	.315
Lewis, John K., Jr , Washington	156	668	210	314

* This column is headed "PC" in the original table
Source *Spalding's Official Baseball Guide 1938*, pp 112-113, American Sports Publishing Company, New York.

Batting averages.   The familiar batting average of the sport pages of the daily paper is a ratio of the hits made by a batter in relation to the total number of times he was at bat.   Table 23 shows a series of selected batting averages.   The figures in the last column of Table 23 may be correctly thought of as either ratios to one or as averages of a series of observations

each having a value of 1 or 0 (that is, either the batter did or did not make a hit). If a man has been at bat 75 times and has made 25 hits, his batting average would be shown as .333 and is spoken of as "three hundred and thirty-three." If he had made a hit every time he was at bat, his figure would be 1.000, which is referred to as "one thousand." Notice that certain contradictions are involved in some of the terms used to refer to these data. The column of figures is frequently headed "percentage", the figures are printed as *ratios to one;* the figures are spoken of as *per thousand!*

Airplane accident ratios. The safety of air travel may be reflected by means of ratios. The number of miles flown during a year (or other convenient period) may be divided by the number of accidents to obtain "miles flown per accident." In 1937 domestic air lines flew 66,071,507 miles and 42 accidents occurred. The lines therefore flew 1,573,131 miles per accident. In the same year there were 5 accidents involving a fatality, and dividing the mileage flown by 5 gives 13,214,301 miles per fatal accident. During 1937 there were 40 passenger fatalities as a result of airplane accidents, and it appears that domestic air lines flew 1,651,788 miles per passenger fatality. Passenger fatalities may be related to passenger miles, and since domestic air lines flew 476,603,165 passenger miles in 1937 we have $476,603,165 \div 40 = 11,915,079$ passenger miles flown per passenger fatality. Because of the small number of accidents and fatalities involved, these ratios fluctuate tremendously from year to year. For example, the passenger miles flown per passenger fatality were 3,500,607 in 1930; 21,686,515 in 1933; 11,050,508 in 1934; 20,927,034 in 1935; and 9,903,188 in 1936. It will be observed that, as air travel becomes safer, all of the ratios mentioned will grow larger. It would also be possible (though not customary) to compute the ratio of the number of accidents or fatalities per million miles flown. Such ratios would be reciprocals of those given and, as air travel becomes safer, would approach zero.

The 100 per cent statement. When banks, insurance companies, and other corporations present financial information to the public, they find it effective to supplement the dollar figures with percentages. Thus a financial statement may show each asset as a percentage of all assets, and each liability as a percentage of all liabilities. The procedure is particularly effective when the dollar figures are large. Table 24 shows the assets of the New York Life Insurance Company as set forth in an annual report. The actual figures are too large for the ordinary reader to grasp and compare, but the percentage data make comparisons less difficult. In preparing such a percentage statement it is desirable not to show too many decimal places, else comparisons cannot readily be made. A recent statement of the resources of a bank carried all percentages to three decimal places. This was quite unnecessary, particularly since the smallest item

"sundry securities" was .035 (.0349) per cent and could have been shown as .03 per cent, and since the second smallest item "other assets" was .039 per cent and could have been shown as .04 per cent. For popular presentation there is some advantage in lumping such small items together in order to center attention upon the more important ones. These two small items, if combined, would have appeared as .07 per cent or as .1 per cent, and all percentages could have been shown to but one decimal

TABLE 24

ASSETS OF THE NEW YORK LIFE INSURANCE COMPANY, DECEMBER 31, 1937

Asset	Amount*	Per cent of total
Cash on hand, or in bank .	$ 64,231,858.43	2.55
United States Government, direct, or fully guaranteed bonds . ..	512,300,999.54	20.33
State, county and municipal bonds .	254,845,789 65	10.11
Railroad bonds	297,213,924.28	11 79
Public utility bonds	229,437,611.57	9.10
Industrial and other bonds	49,549,133.97	1 97
Canadian bonds .	59,771,724.10	2.37
Foreign bonds . . .	133,671.00	01
Preferred and guaranteed stocks .	81,644,201 00	3 24
Real estate owned (including home office)	140,089,034 62	5 56
Foreclosed real estate subject to redemption	2,265,334.31	.09
First mortgages on city properties	405,082,891.33	16.07
First mortgages on farms .	6,936,336.77	.27
Policy loans	355,265,818.60	14 09
Interest and rents due and accrued .	30,149,211.77	1 20
Net amount of uncollected and deferred premiums .	31,358,413.78	1.24
Other assets	74,261 64	.01
Total admitted assets . .	$2,520,350,216 36	100.00

* Bonds eligible for amortization are carried at their amortized values determined in accordance with the laws of the state of New York All other bonds and all guaranteed and preferred stocks are carried at market ~~~ ~~ ~~ ~ ~~ ~~~~ by the National Association of Insurance Commissioners Securities amounting ~, ~,~ , ~ ~~ ~ cluded above, are deposited as required by law
  Source .~ · . · *Report of the New York Life Insurance Company,* p 6

place. However, it may have been desired to emphasize the smallness of either "sundry securities" or "other assets," or both.

Railroad ratios. The efficient operation of railroads necessitates the collection and use of a vast amount of statistical data in connection with which numerous ratios are calculated.

The investment per mile of line is obtained by dividing total investment (including roadway, tracks, equipment, stations, shops, etc.) by the number of miles of railroad line. This figure was $105,922 per mile in 1936.

Freight revenue per ton mile is obtained by dividing total freight revenue by the total number of ton miles of freight hauled. In 1937 the "revenue

per ton mile" for class I railroads was .935 cents. Similarly, we may compute the "revenue per passenger mile," which amounted to 1.79 cents in 1937.

The operating ratio is the ratio of operating expenses to operating revenues. In 1937 operating expenses were \$3,119,064,323, while operating revenues were \$4,166,068,600. The operating ratio was 74 87 per cent.

There are a number of other railroad ratios; the meaning of each is rather obvious. Enumerating a few for class I railroads in 1937: the revenue per ton of freight was \$1.85; the haul per ton of freight was 198.0 miles; the revenue per passenger was \$1.59; the haul per passenger was 49 6 miles; the rate of return on aggregate property investment was 2.26 per cent, the hours worked during the year per railroad employee were 2,512, the percentage of unserviceable freight locomotives averaged 25 5 during the year, while 10.1 per cent of the freight cars were in the same condition; the ton miles per day per freight car were 562; the mileage per day per freight car was 32 9 miles [5]

The railroad ratios mentioned above are one type of business ratios. Many sorts of business organizations compute diverse ratios for the better functioning of the enterprise. Discussed in another volume[6] are such ratios as current ratio (current assets ÷ current liabilities), merchandise turnover (net sales ÷ merchandise inventory), margin of profit (profit ÷ sales), labor turnover (replacements ÷ number on payroll), and others.

Faulty Use of Percentages

Ratios and percentages are in such general use that it is not surprising to find them occasionally misused. Difficulties encountered in the calculation and use of percentages can generally be traced to one of the following causes: (1) confusion in regard to the base, (2) calculation of percentages based on small absolute numbers, (3) misplaced decimal points, (4) arithmetic mistakes, (5) improper procedure in averaging percentages, (6) the use of percentages which are awkwardly large. These will be discussed in order.

Confusion in regard to base. Some years ago the dean of a mid-western veterinary college was reported to have stated that over a period of five years (1916 to 1921) the enrollment in veterinary colleges in the United States had decreased 500 per cent. This would indeed mean a very small registration, since a decrease of 500 per cent would mean a negative figure four times the size of the original enrollment! The absolute figures showed

---

[5] For these and other railroad ratios, see *A Yearbook of Railroad Information*, issued annually by the Committee on Public Relations of the Eastern Railroads, New York.

[6] See F. E. Croxton and D. J. Cowden, *Practical Business Statistics*, pp. 139-149 Prentice-Hall, Inc., New York, 1934.

an original enrollment of 3,160 students, which decreased to 641 five years later. The decrease was 2,519 students or 79.7 per cent. What the probably-misquoted dean most likely said was that the enrollment in the earlier year was 500 per cent of the enrollment in the later year.

In the autumn of 1920 a determined effort was made by the United States district attorney to have restaurants in Pittsburgh lower their prices to a pre-war level. Newspapers announcing the success of the drive stated that Pittsburgh restaurants had cut their prices 50 to 100 per cent. It is, of course, clear that prices cannot be cut 100 per cent, else the servings formerly sold would be given away! The price reductions on a number of dishes were stated, the greatest reduction took place in the price of doughnuts and pie. These had formerly sold at 15 cents per order. Identical sized servings were sold at 5 cents after the reduction; hence the reduction amounted to 66.7 per cent of the former selling price.

Accounts appearing in newspapers in the spring of 1934 made public the results of a study of maternal mortality. There had been enumerated 1,343 cases of preventable deaths. The attending physicians were held responsible for 61.1 per cent, the patients themselves were said to be responsible for 36.7 per cent, while midwives were allegedly responsible for 2.2 per cent. The percentages were misleading because doctors attend a large proportion of confinement cases while midwives attend but a few. A proper procedure for calculating the percentages consists of, first, relating the number of deaths for which the physicians allegedly were responsible to the total number of cases attended by physicians, and, second, relating the number of deaths for which midwives presumably were responsible to the total number of cases attended by midwives. Upon such a basis the New York Obstetrical Society announced[7] that, according to the same study mentioned above, "responsibility is ascribed to the physician in 678 maternal deaths, which is 47 per cent of the deaths occurring among patients attended by the physician, while the midwife was responsible for 29 maternal deaths, or 60.4 per cent of the deaths in [among] women attended by the midwife." It was further quite properly pointed out that "it is an almost impossible task to ascribe responsibility in a large percentage of cases. . . ."

Percentages from small numbers. An almost classic illustration of the undesirability of using percentages based upon small numbers is given by Chaddock.[8]

A short time after Johns Hopkins University had opened certain courses in the University to women, it was reported that 33⅓ per cent of the

[7] See the New York *Times*, April 12, 1934.
[8] Robert E. Chaddock, *Principles and Methods of Statistics*, Houghton Mifflin Co., Boston, 1925, pp. 13–14.

women students had married into the faculty of the institution.   Of course the important information was the number of women students.   There were only three.   *When dealing with a small number of cases, the use of percentages alone leads to wrong impressions.*   In these cases either percentages should not be used at all or the numbers upon which they are based should accompany the percentages.

Misplaced decimal points. Mistakes involving misplaced decimal points may lead to gross misinterpretations.   They are a common sort of mistake and should be guarded against.   Sir Josiah Stamp[9] gives a rather unusual illustration:

> A periodical return of revenue received into the Exchequer was laid before Lord Randolph, and his private secretary, Mr. George Gleadowe of the Treasury, was looking over his shoulder, and Lord Randolph expressed satisfaction at the fact that the Customs revenue had increased by 34 per cent. as compared with the corresponding period in the preceding year.   Mr. Gleadowe pointed out to him that it was only .34 per cent.   "What difference does that make?" asked Lord Randolph.   When it was explained to him he said, "I have often seen those damned little dots before, but I never knew until now what they meant."

Another example of a misplaced decimal point is mentioned on page 162.

Arithmetic mistakes. A speaker discussing the upturn in employment in May 1934 pointed out that employment had increased by 32 per cent during the year.   He then said, according to a newspaper, "Cold figures and percentages don't mean very much to one not accustomed to deal with them, but these figures simply say that, as to any group of two men who were employed a year ago, one extra man has been added."   The proper statement is that, as to any group of *three* men who were employed a year ago, one extra man had been added.

Improper averaging of percentages. The occasional necessity for averaging percentages calls for mention of a pitfall and for consideration of the proper procedure.   Consider the following figures:

FOREIGN-BORN POPULATION OF THE NEW ENGLAND STATES, 1930

State	Total white population	Foreign-born white population	Per cent foreign-born
Maine	795,183	100,368	12.6
New Hampshire	464,350	82,660	17.8
Vermont	358,965	43,061	12.0
Massachusetts	4,192,926	1,054,636	25.2
Rhode Island	677,016	170,714	25.2
Connecticut	1,576,673	382,871	24.3

[9] Sir Josiah Stamp, *Some Economic Factors in Modern Life*, p. 265.   P. S. King and Son, London, 1929

It is desired to know the average proportion of foreign-born white persons for the New England division. If we add the six percentages and divide by six, we have $117.1 \div 6 = 19.5$ per cent. This figure, however, does not correctly represent the situation; the six percentages were calculated from different bases and therefore should be weighted accordingly. The easiest procedure for obtaining the correct percentage consists of totaling the white population for the six states (8,065,113 persons), totaling the foreign-born white population (1,834,310 persons), and dividing the second figure by the first. The result is 22.7 per cent, which is the proportion of foreign-born white persons in the New England division. The same result could also be obtained by averaging the six percentage figures, provided each is weighted according to the base from which it has been calculated. This procedure of multiplying each percentage by its base, summing the results, and dividing by the sum of the base figures (or weights) is essentially the same as the method just used. The result, however, is a little less accurate since each percentage figure has been rounded. The error involved in rounding a given percentage is magnified when the percentage is multiplied. But since some percentages are understated and some are overstated, there is a *tendency* for these errors to counterbalance.

Unduly large percentages. While percentages are extremely useful for purposes of comparison, percentage figures which are very large may serve to confuse. The New York *Times* for July 27, 1933, states in a headline "N. Y. Central Gains 2000% for Month." The net operating income for the New York Central Railroad was $192,052 for June 1932 and $4,384,965 for June 1933. The latter figure is 2183 per cent greater than the former. The use of a figure as great as 2000 per cent is meaningless to many people, and a more accurate impression could probably be conveyed by indicating that June 1933 operating income was 22 times larger than June 1932 operating income; or, it would also be correct to say that June 1933 operating income was 23 times as great as June 1932 operating income.

That the use of large percentage figures may lead to difficulties is vividly brought out in the same paper for April 4, 1935. The statement is made that a certain bank "grew by 3000 per cent." The bank referred to was the Union Trust Company of Pittsburgh, the resources of which mounted from $100,000 to $300,000,000. The second figure is actually *3000 times* the first figure, or, in percentages, *300,000 per cent* of the smaller figure. The *growth* then was 299,900 per cent, or 2999-fold.

Selected References

W. B. Bailey and J. Cummings· *Statistics*, Chapter V[; A. C. McClurg and Co., Chicago, 1917. Basic principles.

F. E. Croxton and D. J. Cowden *Practical Business Statistics*, Chapter VII; Prentice-Hall, Inc., New York, 1934. Includes a treatment of selected business ratios.

Harry Jerome: *Statistical Method*, Chapter VIII; Harper and Brothers, New York, 1924.

G. C. Whipple· *Vital Statistics, an Introduction to the Study of Demography* (Second Edition), Chapters IV, VII, VIII, XI, John Wiley and Sons, Inc , New York 1923.   Probably the best discussion of birth rates and death rates.

CHAPTER VIII

THE FREQUENCY DISTRIBUTION

One method of organizing and summarizing statistical data consists in the formation of a frequency distribution. In this device the various items of a series are classified into groups and the number of items falling into each group is stated. A frequency distribution is shown in Table 27. Sometimes the user of statistics will find frequency distributions already constructed in the publications to which he may refer; sometimes he will construct his own frequency distribution from unclassified data. We shall begin our discussion of the frequency distribution by first considering the appearance of the raw or unclassified data.

Raw Data

The unclassified data from which a frequency distribution might be made may appear as do the data of Table 25. Here we have the grades made by the 1937 graduating class of the United States Naval Academy for the 4-year course. The arrangement of the grades is according to the alphabetical order of the midshipmen's names, though we have omitted the names in order to save space. Another illustration of raw data, from which a frequency distribution might be constructed, is the payroll of a factory. The employees on the payroll may be listed alphabetically by name; by employee number; by departments, and then by name or number; by seniority; or in some other convenient order. Considering the grades of the midshipmen as shown in Table 25, it is apparent that very little information is forthcoming unless the figures are rearranged. When the data are in this form, it is a tedious task to find even the lowest grade and the highest grade. It is yet more difficult to ascertain around what value the grades tend to concentrate, or if indeed they do show such a concentration. These and other steps in analysis are facilitated by rearranging and summarizing the data.

164

TABLE 25

Grades Received in the 4-Year Course by Members of the 1937 Graduating Class of the United States Naval Academy

(Alphabetical listing, names omitted for brevity)

78 9	78 2	71 2	74 8	74 6	74.4	85.9	74.2	84.7	77.9
80 7	71 3	79 2	86 3	76.1	72 1	76.3	72.1	79.2	74.6
80 2	78.7	78 2	74 0	72 9	79 0	70 7	74 4	78 5	74.1
76 5	73.6	79 3	73 4	75 5	75 8	72 8	73.5	86 2	76.5
73 5	75 1	79 2	78 0	81 2	71 1	81 1	77.7	78 1	79.6
89 5	76 5	75 4	75 5	81 6	79 2	92 1	85.4	77 4	91.0
79 9	81 5	69 2	85 1	76 7	76 1	74 4	78.8	74 9	85 0
74 9	76 2	81 5	76 3	74 5	82 5	80 8	70.6	74 8	73.6
74 2	73 7	83 2	74 9	75 7	73 4	75 2	70.7	79.5	77 8
89 2	73 3	81 0	90 2	73.0	82 7	76 0	72.2	80 4	77 2
86 4	73 8	71 4	74 7	79 9	80 3	73 3	74.9	78 2	79.8
77 2	81 2	80 4	78 2	77 8	77.6	75 1	73.5	74 8	77 0
80 4	73 9	69.8	82 3	83 4	79 3	81 9	84.4	81.3	72.6
73 2	78 6	76 2	79.7	87 3	84 2	75 5	78.3	79 1	75 3
74 3	84 3	77 8	81 9	81 2	83 3	89 7	85.8	74 2	79 2
82 4	80 8	71 1	75 0	83 7	75 4	78 6	76.8	76 0	74 1
80 4	73.4	71 5	77 4	85 7	76 0	79 0	75.2	77 6	76 5
76.7	76 0	76 2	82 1	82 9	79 0	74 4	75 5	77.3	82 1
76 0	74 5	77 1	79 7	73 0	81 3	77 4	77 6	79 4	81 1
84 2	84.3	72 4	77 6	82 1	72 1	79 1	74 6	71 7	86 5
81 3	80 3	77 8	77 2	78.8	74 8	74 4	76 4	72 9	72 0
75 3	73 7	82 7	82 0	86 8	78 2	77.6	71.8	71.2	73 8
72.3	77.5	71 3	86 5	80 6	86.1	74 2	75 6	76 6	74 0
79.3	71 9	81 9	84 7	73 9	79 1	71 7	78.6	84 5	89 1
74 9	77 5	73 7	72 3	78.0	78 2	77 2	80.4	86.3	74 4
76.3	77 5	83 9	79 7	76.2	81.0	74.9	84 5	83.5	73 5
74 6	75.1	79.1	78 5	82.0	75 4	82 2	73 5	76.4	68.8
86 1	74 4	75.1	71 9	81.5	81 9	73 8	81.1	86 2	77 9
78 7	68 9	78 2	78.9	77.8	78.5	81.0	80 4	78.7	74 5
76 4	80 1	72 9	75 4	72 8	87 0	80 1	77 5	75.2	83.3
75 7	77 4	74.5	82.8	75.9	76.4	77 3	74 4	83.4	
71 4	79 6	74 4	72.6	79.8	77.2	73 2	85.0	78.3	
85 2	76 6	78 6	75 1	85 4	76 4	86 7	75 7	83.0	

Source Adapted from *Annual Register of the United States Naval Academy, 1937-1938*, pp 41–46, which gives grades in terms of a maximum of 1000

The Array

In Table 26 the midshipmen's grades have been rearranged in descending order. Such an arrangement (whether ascending or descending) is called an *array*. It arranges the items in order of magnitude. We have not summarized; that will be done when we construct the frequency distribution. A consideration of the array puts us in a position to learn something from the data. First, the array enables us to see at once the range of the grades, which varied from 68.8 to 92.1. Second, it may also be seen that there is a concentration of grades somewhere between 74 and 78. This

will be more clearly seen when we examine the frequency distribution and consider measures of central tendency Third, a somewhat more extended examination gives us a rough idea of the distribution of the grades. We

TABLE 26

ARRAY OF GRADES RECEIVED IN THE 4-YEAR COURSE BY MEMBERS OF THE 1937 GRADUATING CLASS OF THE UNITED STATES NAVAL ACADEMY

92 1	84.4	81 5	79 7	78 5	77 4	76 1	74 9	74 1	72.6
91 0	84 3	81 5	79 7	78 5	77 4	76 1	74 9	74 0	72.4
90 2	84 3	81 3	79 7	78 5	77 4	76 0	74.9	74 0	72.3
89 7	84 2	81 3	79 6	78 3	77 3	76 0	74 9	73 9	72 3
89 5	84 2	81 3	79 6	78 3	77 3	76 0	74.9	73 9	72.2
89 2	83 9	81 2	79 5	78 2	77 2	76 0	74 8	73 8	72.1
89 1	83.7	81 2	79.4	78 2	77 2	76 0	74 8	73 8	72.1
87 3	83 5	81.2	79 3	78.2	77 2	75 9	74 8	73 8	72 1
87 0	83.4	81 1	79 3	78 2	77.2	75 8	74.8	73 7	72.0
86 8	83.4	81 1	79 3	78 2	77 2	75.7	74 7	73.7	71.9
86.7	83 3	81.1	79 2	78 2	77 1	75 7	74.6	73 7	71.9
86.5	83.3	81.0	79 2	78 2	77 0	75.7	74 6	73 6	71 8
86 5	83.2	81 0	79 2	78 1	76 8	75 6	74 6	73 6	71 7
86 4	83.0	81 0	79 2	78 0	76 7	75.5	74 6	73.5	71.7
86.3	82 9	80 8	79 2	78 0	76.7	75 5	74 5	73 5	71 5
86.3	82.8	80 8	79 1	77 9	76 6	75 5	74.5	73.5	71 4
86 2	82.7	80 7	79 1	77 9	76 6	75.5	74.5	73 5	71.4
86 2	82.7	80 6	79 1	77 8	76 5	75 4	74 5	73 5	71.3
86.1	82 5	80.4	79 1	77 8	76 5	75 4	74.4	73 4	71 3
86.1	82.4	80 4	79 0	77 8	76.5	75.4	74 4	73.4	71 2
85 9	82.3	80 4	79 0	77 8	76 5	75 4	74.4	73.4	71.2
85 8	82.2	80 4	79 0	77 8	76 4	75.3	74 4	73 3	71.1
85 7	82.1	80.4	78.9	77.7	76.4	75.3	74.4	73 3	71.1
85 4	82.1	80.4	78 9	77.6	76 4	75 2	74 4	73 2	70 7
85 4	82.1	80 3	78 8	77.6	76 4	75 2	74 4	73 2	70 7
85.2	82.0	80 3	78 8	77 6	76 4	75.2	74 4	73.0	70 6
85 1	82.0	80 2	78 7	77.6	76 3	75 1	74 4	73 0	69 8
85.0	81 9	80 1	78 7	77 6	76 3	75.1	74.3	72 9	69.2
85.0	81.9	80.1	78 7	77.5	76 3	75.1	74.2	72.9	68.9
84.7	81.9	79 9	78.6	77.5	76 2	75.1	74.2	72 9	68.8
84.7	81.9	79 9	78.6	77 5	76.2	75 1	74.2	72 8	
84 5	81.6	79 8	78 6	77 5	76 2	75.0	74.2	72.8	
84.5	81.5	79.8	78 6	77 4	76.2	74 9	74.1	72.6	

may observe, for example, that there are few grades below 71 or above 86. This particular feature of the series will be much more readily studied when we have the frequency distribution. Fourth, it may be noticed that the figures show a fairly regular continuous change. If the grades are expressed as whole percentages, all consecutive values from 69 to 87 are represented. If we consider the figures as shown, to one decimal place, we may observe only three values not represented from 73.2 to 80.4, within which range 205 of the 327 midshipmen occur. If the grades had been for a

larger number of students, this tendency would have been more marked.

The array, however, is a cumbersome form of the data.   Furthermore, it is troublesome to construct, because of the necessity of rearranging all the items.   One fairly satisfactory method of constructing an array consists of recording the figures on small cards and sorting the cards.   Of course, if the data are punched on mechanical tabulating cards, the construction of an array is simple.

When studying grades, we may frequently want to make an array. The Naval Academy publishes each year a Merit Roll of the graduating class, listing the names and standings of the midshipmen in the order which

TABLE 27

FREQUENCY DISTRIBUTION OF
GRADES OF THE 1937 GRADUATING
CLASS OF THE UNITED STATES
NAVAL ACADEMY

Grade	Number of midshipmen
68 0–69.9	4
70 0–71.9	17
72 0–73 9	39
74 0–75 9	62
76 0–77.9	58
78 0–79 9	52
80 0–81 9	35
82 0–83.9	22
84 0–85 9	18
86 0–87 9	13
88 0–89 9	4
90 0–91.9	2
92 0–93 9	1
Total	327

is given in Table 26.   If we are interested in a campaign to raise funds for a hospital or community chest, it might be very useful (for publicity purposes, for example) to list the individual gifts in descending order.   It is obvious, however, that such a listing of 500 or 1,000 contributions would be cumbersome and of limited value.   In many instances there is no particular advantage in making an array.   It would be a waste of time for a concern to make an array of the amounts paid to its employees each month.   There is not much reason why a bank should make an array of the daily balances of its many depositors.   On the other hand, a student of vital statistics might find it very valuable in a study of birth rates to array the various cities in ascending or descending order and consider the reasons for the differences.

The Frequency Distribution

The array of Table 26 rearranged the midshipmen's grades. The frequency distribution of Table 27 summarizes the grades into 13 groups or classes. It is obvious that the frequency distribution does not show the details given in the array, but much is gained by the summarization. We can see that the lowest grade is not below 68 and that the highest grade is not quite 94; we cannot ascertain the exact values of the highest and lowest grades as we did from the array. The concentration of grades in the neighborhood of 74–78 is apparent at a glance. If we draw a curve of the frequency distribution, as in Chart 81, we can visualize the data readily and we may make comparisons with other series as discussed in later sections

NUMBER OF
MIDSHIPMEN
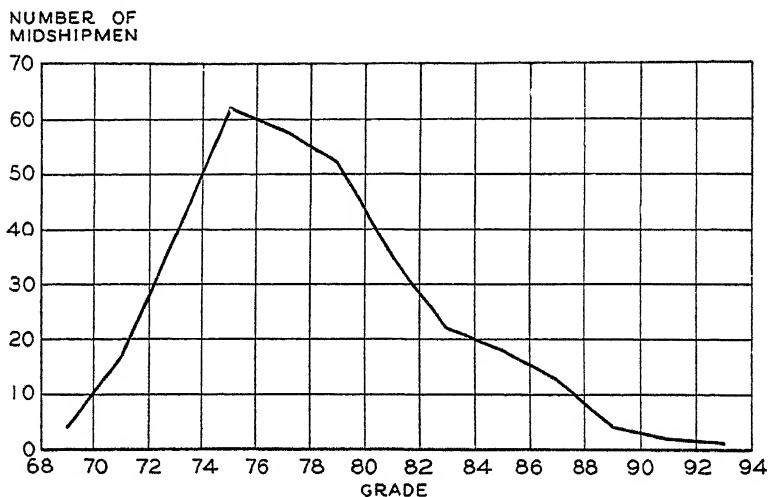
Chart 81. Grades of the 1937 Graduating Class of the United States Naval Academy.
(Data of Table 27.)

of this chapter. Having classified the data, we are in a position to make rapid computations of certain values (discussed in the following chapters) which will assist us in describing and analyzing the data.

When an array is available, the frequency distribution may be made by merely counting the items. It is not advisable, however, to make an array solely for the purpose of making the frequency distribution, because too great an amount of time is required to construct the array.

If the data are in unorganized form, as in Table 25, we may construct a frequency distribution by a scoring device similar to that shown in Chapter II. Another method of handling the figures consists of making an entry form such as that of Table 28. This is less laborious than making an array

TABLE 23

Entry Form for Grades of the 1937 Graduating Class of the United States Naval Academy

68 0–69 9	70 0–71 9	72 0–73.9	74 0–75 9	76 0–77 9	78 0–79 9	80 0–81 9	82 0–83 9	84 0–85 9	86 0–87.9	88.0–89.9	90.0–91.9	92.0–93.9
68 9	71 4	73 5	74 9	76 5	78 9	80 7	82 4	84.2	86.4	89 5	90 2	92.1
69 2	71 3	73 2	74 2	77 2	79 9	80 2	83 2	85 2	86.1	89.2	91.0	
69 8	71 9	72 3	74 3	76 7	79.3	80 4	82 7	84 3	86 3	89.7		
68 8	71 2	73 6	75 3	76 0	78 7	80 4	83 9	84 3	86 5	89.1		
	71 4	73 7	74 9	76 3	78 2	81 3	82 3	85 1	87 3			
	71 1	73 3	74 6	76 4	78 7	81 5	82 1	84 7	86 8			
	71 5	73 8	75 7	76 5	78 6	81 2	82 0	85 7	86 1			
	71 3	73.9	75 1	76 2	79 6	80 8	82 8	85 4	87.0			
	71 9	73 4	74 5	76 0	79 2	80 3	83 4	84.2	86.7			
	71 1	73 7	75 1	77 5	78 2	80 1	83 7	85 9	86.2			
	70 7	72 4	74 4	77 5	79 3	81 5	82 9	85 4	86.3			
	71 7	73.7	75 4	77.5	79 2	81 0	82 1	84 4	86.2			
	70.6	72 9	75 1	77 4	79 1	80 4	82 0	85 8	86 5			
	70 7	73 4	74 5	76 6	78 2	81 9	82 5	86 5				
	71 8	72 3	74 4	76 2	78 6	81 9	82 7	84 5				
	71 7	72 6	74 8	77 8	78 0	81 2	83 3	85 0				
	71 2	72 9	74.0	76 2	78 2	81 6	82 2	84 7				
		73 0	75 5	77.1	79 7	81 2	83 5	84 5				
		73 0	74 9	77 8	79 7	80 6	83 4	85 0				
		73 9	74 7	76 3	79 7	81 5	83 0					
		72.8	75 0	77 4	78 5	80 3	82 1					
		72 1	75.4	77 6	78 9	81 3	83 3					
		73 4	75 1	77 2	79 9	81 0						
		72 1	74 6	76 1	78 8	81 9						
		72 8	75 5	76 7	78 0	81 1						
		73 3	74 5	77 8	79 8	80 8						
		73 8	75 7	76 2	79.0	81 9						
		73 2	75 9	77 8	79 2	81 0						
		72.1	74 4	76 1	79 3	80.1						
		73 5	75 8	77.6	79 0	80 4						
		72 2	75 4	76 0	78 2	81 1						
		73 5	74 8	76 4	79.1	80 4						
		73 5	75.4	77 2	78.2	80 4						
		72 9	74 4	76 4	78 5	81 3						
		73 6	75 2	76.3	78 6	81 1						
		72 6	75 1	76 0	79 0							
		72 0	75 5	77 4	79 1							
		73 8	74 4	77 6	78 8							
		73.5	74 4	77 2	78 3							
			74 2	77 3	78 6							
			74 9	77 7	79 2							
			74 2	76 8	78 5							
			74 4	77.6	78 1							
			74.9	76 4	79 5							
			75 2	77 5	78 2							
			75 5	77 4	79.1							
			74 6	76.0	79.4							
			75 6	77 6	78.7							
			74 4	77 3	78 3							
			75 7	76.6	79 6							
			74.9	76 4	79 8							
			74.8	77 9	79 2							
			74 8	76 5								
			74.2	77 8								
			75 2	77 2								
			74.6	77.0								
			74.1	76 5								
			75.3	77 9								
			74.1									
			74.0									
			74.4									
			74.5									

and has certain advantages over the scoring procedure. The advantages
of the entry form are: (1) We can scan the columns to see if any item is
incorrectly entered. (2) We can total the items entered and check this
total against the total of the unclassified data. (3) If we should decide
that we want classes of 1 per cent or 3 per cent instead of 2 per cent, we

TABLE 29

AVERAGE HOURLY EARNINGS OF 13,427 WAGE EARNERS IN OPEN-HEARTH FURNACES
1935

Average hourly earnings (cents)	Number of wage earners	Frequency densities, number of wage earners per 2 5 cents of earnings
25 0 and under 27 5	1	1
27 5 and under 30 0	27	27
30 0 and under 32 5	13	13
32 5 and under 35.0	43	43
35 0 and under 37.5	90	90
37 5 and under 40 0	98	98
40 0 and under 42.5	289	289
42 5 and under 45.0	360	360
45 0 and under 47 5	779	779
47 5 and under 50 0	1,284	1,284
50.0 and under 55 0	1,437	718 50
55 0 and under 60.0	1,263	631 50
60 0 and under 65 0	1,134	567 00
65.0 and under 70 0	1,081	540 50
70 0 and under 75.0	957	478 50
75 0 and under 80 0	777	388.50
80 0 and under 85.0	613	306.50
85 0 and under 90 0	577	288 50
90.0 and under 100 0	809	202 25
100.0 and under 110 0	546	136 50
110.0 and under 120 0	287	71.75
120 0 and under 130 0	319	79 75
130.0 and under 140 0	202	50 50
140 0 and under 150 0	129	32 25
150 0 and under 160 0	83	20 75
160 0 and under 170 0	82	20 50
170.0 and over	147	.
Total .	13,427	

Source *Monthly Labor Review*, Vol 42, No 4 (April 1936), p 1045

can re-form our frequency distribution with little effort. (4) As will be
shown in the next chapter, the entry form enables us to find out how closely
the mid-value of a class agrees with the average of the items in that class

If desired, the classes used in the entry form may be narrower than we think we shall want for the frequency distribution. These classes may then be readily combined into wider ones, using whatever interval and whatever class limits seem advisable.

All the class intervals of the frequency distribution of Table 27 are 2 per cent. Charting and computations are facilitated when the class inter-

NUMBER
OF TIRES

Chart 82. Mileage of 98 Automobile Tires, Size 4.75 x 19; 1,000-Mile Class Intervals. (Data from a confidential source The tires were used by a fleet of delivery trucks in and around New York City)

vals are all the same. Whenever possible, therefore, frequency distributions should be constructed with uniform class intervals. This, however, is not always practicable (see for example Table 42). In Table 29 there is also shown a frequency distribution which has non-uniform class intervals. In this instance the result is to give more detailed information for the groups having lower earnings.

Selecting the number of classes. No hard and fast rule can be given as to the number of classes into which a frequency distribution should be divided. If there are too many classes, many of them will contain only a few frequencies and the distribution will show noticeable irregularity when plotted. If there are too few classes, so many frequencies will be crowded into a class as to cause much information to be lost. Chart 82 shows a series which is rather irregular because the mileage records of 98 automobile tires were grouped into 32 classes each 1,000 miles in width. The curve assumes a much more regular outline in Chart 83 when the same data are put into 12 classes of 3,000 miles each. The number of classes to use depends upon the number of frequencies in the series and the

and has certain advantages over the scoring procedure.   The advantages of the entry form are· (1) We can scan the columns to see if any item is incorrectly entered.   (2) We can total the items entered and check this total against the total of the unclassified data.   (3) If we should decide that we want classes of 1 per cent or 3 per cent instead of 2 per cent, we

TABLE 29

Average Hourly Earnings of 13,427 Wage Earners in Open-Hearth Furnaces
1935

Average hourly earnings (cents)	Number of wage earners	Frequency densities, number of wage earners per 2 5 cents of earnings
25 0 and under 27 5	1	1
27 5 and under 30 0	27	27
30 0 and under 32 5	13	13
32 5 and under 35 0	43	43
35 0 and under 37 5	90	90
37 5 and under 40 0	98	98
40 0 and under 42 5	289	289
42 5 and under 45 0	360	360
45 0 and under 47 5	779	779
47 5 and under 50 0	1,284	1,284
50 0 and under 55 0	1,437	718 50
55 0 and under 60 0	1,263	631 50
60 0 and under 65.0	1,134	567 00
65 0 and under 70 0	1,081	540.50
70 0 and under 75.0	957	478 50
75 0 and under 80 0	777	388 50
80.0 and under 85 0	613	306 50
85.0 and under 90 0	577	288 50
90 0 and under 100 0	809	202.25
100.0 and under 110 0	546	136 50
110.0 and under 120 0	287	71 75
120 0 and under 130 0	319	79 75
130 0 and under 140 0	202	50 50
140.0 and under 150 0	129	32 25
150 0 and under 160 0	83	20 75
160 0 and under 170.0	82	20 50
170.0 and over	147	
Total . .	13,427	

Source *Monthly Labor Review*, Vol 42, No 4 (April 1936), p 1045

can re-form our frequency distribution with little effort.   (4) As will be shown in the next chapter, the entry form enables us to find out how closely the mid-value of a class agrees with the average of the items in that class

If desired, the classes used in the entry form may be narrower than we think we shall want for the frequency distribution. These classes may then be readily combined into wider ones, using whatever interval and whatever class limits seem advisable.

All the class intervals of the frequency distribution of Table 27 are 2 per cent. Charting and computations are facilitated when the class inter-

Chart 82. Mileage of 98 Automobile Tires, Size 4.75 x 19; 1,000-Mile Class Intervals. (Data from a confidential source The tires were used by a fleet of delivery trucks in and around New York City)

vals are all the same. Whenever possible, therefore, frequency distributions should be constructed with uniform class intervals. This, however, is not always practicable (see for example Table 42). In Table 29 there is also shown a frequency distribution which has non-uniform class intervals In this instance the result is to give more detailed information for the groups having lower earnings.

Selecting the number of classes. No hard and fast rule can be given as to the number of classes into which a frequency distribution should be divided. If there are too many classes, many of them will contain only a few frequencies and the distribution will show noticeable irregularity when plotted. If there are too few classes, so many frequencies will be crowded into a class as to cause much information to be lost. Chart 82 shows a series which is rather irregular because the mileage records of 98 automobile tires were grouped into 32 classes each 1,000 miles in width. The curve assumes a much more regular outline in Chart 83 when the same data are put into 12 classes of 3,000 miles each. The number of classes to use depends upon the number of frequencies in the series and the

regularity with which the frequencies are distributed within the range of values. The greater the number of frequencies, the more classes we may have. Also, the more regular the distribution of the frequencies, the more classes we may use, since data having a high degree of regularity may be divided into a large number of classes without showing gaps and irregularities in the frequencies. In general it might be said that fewer than 6 or 8 classes should rarely be used, and that more than 16 classes would be useful only for working with extensive data. When the number of classes
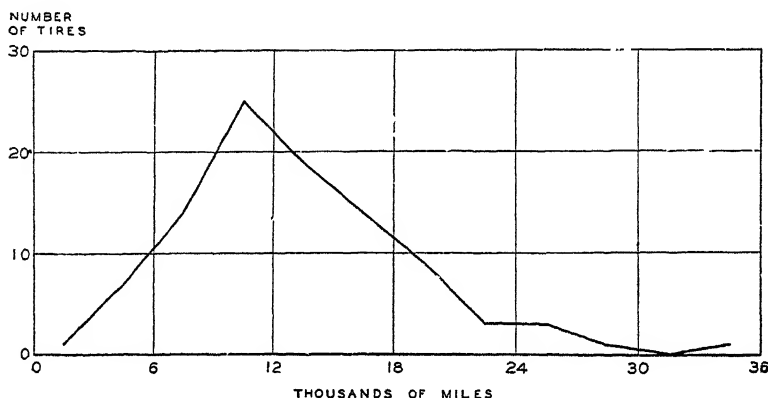
Chart 83. Mileage of 98 Automobile Tires, Size 4.75 x 19; 3,000-Mile Class Intervals. (The vertical scale for this chart is one-third that of Chart 82 in order to compensate for the fact that the class intervals used for this chart are three times those of the preceding chart For source of data see Chart 82)

has been determined, the range of values for the entire distribution indicates the class interval to be used.

Selecting class limits. It was pointed out in Chapter IV that the mid-value of each class is taken to be representative of the class. The mid-values of the classes are made use of not only when charting the frequency distribution, but also in making various computations to be discussed in later chapters. If the limits of each class are not clearly indicated, the mid-value, which is the average of the upper and lower limits, cannot be properly determined. The adequacy of the mid-value assumption will be discussed more fully in Chapter IX. It is important at this point to make clear that, when a frequency distribution is being constructed, the class limits should be so chosen that the mid-value of each class will coincide, so far as possible, with any values around which the data tend to be concentrated.

Suppose that measurements are made of the academic standing of a large group of college freshmen upon a numerical scale ranging from 0 to

100. The data could be expected to be graduated smoothly from, say, 30 to nearly 100. There would be students rating 88.0 and others 89.0; in addition there would be still others falling between these two values. If a large enough group were to be measured the minuteness of the variations between 88.0 and 89.0 would be limited only by the accuracy of the measuring instrument (in this case, the grading system). There would not be a series of values around which the frequencies would tend to concentrate, and the problem mentioned at the end of the preceding paragraph would not arise.

On the other hand, consider the meal checks of a cafeteria, many (but not all) of which are a multiple of 5 cents. In this instance the class intervals should be written 8–12 cents, 13–17 cents, 18–22 cents, etc., thus giving mid-values of 10 cents, 15 cents, 20 cents, etc., which coincide with the concentration points.

The data of freshmen grades and the ratings of midshipmen are illustrations of what is termed a *continuous* variable, since the values are capable of infinitely small variations from each other. Heights and weights of people are also continuous variables. Length of life is another illustration. The data of cafeteria meal checks are illustrative of *discrete* or *discontinuous* data, since the values differ from each other by finite amounts—in this case, one cent. A discrete variable need not show the concentrations which were present in the meal check data. For example, if a large group of workmen are employed at similar tasks and are paid on a piece-rate basis (that is, upon the basis of amount produced), it is quite possible that there may be individuals receiving \$21.21, \$21.22, \$21.23, etc., for a week's work. Although piece rates might be, and often are, in fractions of a cent, the weekly payment must be in terms of whole cents.

The foregoing suggests an important consideration; namely, that we are not so much concerned with the fact that a variable is discrete as we are with the fact that the data may be *broken* and that there are inherent gaps and concentrations in the actual data in hand. The twenty-second annual report (for the year 1935) of the Board of Governors of the Federal Reserve System lists the salaries paid to all of the 328 officers and employees of the Board of Governors. These salaries range from \$840 per annum to \$15,000. There is in no sense an evenly graduated distribution between these limits. The gaps between adjacent values range from \$10 to \$5,000, and there are pronounced concentrations at various customary salaries such as \$1,500, \$1,800, \$2,000, \$2,500, \$3,000, \$3,600, etc. The selection of class limits for a distribution of this type presents great difficulty, as often it is not possible to adjust the mid-values to coincide with all concentration points. An approximate adjustment must then suffice.

The fact that we may be dealing with a continuous variable does not warrant us in selecting class limits blindly. If data are being collected concerning weights of individuals, reported to the *nearest* pound, persons reported as weighing 142 pounds would vary between 141.5 pounds and 142.5 pounds; as a group they would average about 142 pounds. Suppose, however, that weight is reported to the *last* full pound. In that event persons reported as weighing 142 pounds would vary between exactly 142 pounds and just under 143 pounds; as a group they would average about 142.5 pounds. Let us assume that a frequency distribution with class interval of 3 pounds is to be formed. If weights have been reported to the nearest pound, it is correct to write class intervals "142–144, 145–147, 148–150," etc., with mid-values of 143, 146, 149, etc. If, however, weights have been reported to the last full pound, the above is incorrect, but it is correct to write "142 and under 145, 145 and under 148, 148 and under 151," etc., with mid-values of 143.5, 146 5, 149.5, etc.

Two cautions should be noted concerning the writing of class limits. In the first place, designations such as "$50–$100," "$100–$150," etc., should never be used, since the limits overlap. A reader cannot be sure whether $100 belongs in the first class or the second. If tally sheets or entry forms are made with such overlapping class limits, it is possible that $100 items may sometimes be placed in the $50–$100 class and sometimes in the $100–$150 class. If the data are originally given in dollars, the class intervals for both the work sheet and the frequency distribution should read "$50–$99," "$100–$149," etc. If the data are in dollars and cents, the class intervals should read "$50.00–$99.99," "$100.00–$149.99," etc. In rare instances an investigator will write his class limits "$50.01–$100.00," "$100.01–$150.00," etc. The former arrangement is more desirable. The second caution is to avoid writing class limits so that they are mutually exclusive. For example, the wording "over $50 but under $100," "over $100 but under $150," is incorrect because $100 is excluded from both classes.

Curves of frequency distributions. The graphic representation of a frequency distribution was discussed in Chapter IV. Although a frequency distribution may be represented either by a column diagram or a curve, it is usual to employ the latter device. (We shall make use of the column diagram in Chapter XI.) One advantage of the curve is that two or more curves may readily be drawn on the same axes for purposes of comparison. In any event, the first step in the analysis of a frequency distribution should be the construction of a chart, for it will tell us at a glance with which of the following types of distributions we are dealing

Chart 81 shows us the graphic appearance of the data of midshipmen's grades which are shown in Table 27. Although rather regular in appear-

ance, this curve is not symmetrical, but is slightly skewed to the right. (Skewness is discussed in Chapter X.) Many frequency distribution curves encountered in the social sciences are asymmetrical and frequently are skewed to the right. Only rarely do we find a curve skewed to the left.

Biological and anthropometrical series (especially those involving linear measurements, such as height, rather than two- or three-dimension measurements, such as waist circumference or weight), frequently yield curves which are almost symmetrical Witness the curve of the basal diameter of the egg masses of snails, shown as Chart 116. Another roughly symmetrical series is shown in Chart 84, which pictures the height distribution of a large group of male industrial workers.
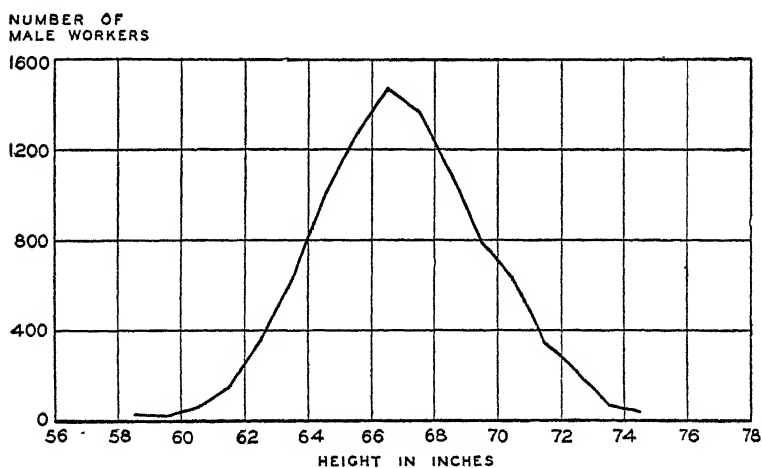
Chart 84.   Heights of 9,552 Male Industrial Workers.   (Data from *A Health Study of Ten Thousand Male Industrial Workers*, p. 59.   United States Public Health Service, Public Health Bulletin No. 162)

A curve which is skewed to the left appears in Chart 85, which shows the batting averages of 881 players of the two major leagues (the American League and the National League) and the three leading minor leagues (the American Association, the International League, and the Pacific Coast League). These 881 players include all those who had played in 10 games or more and who had been at bat at least 25 times. Thus the figures include both substitutes and regular players. If the substitutes are excluded from the data, the resulting curve for regular players approaches symmetry (see Chart 119). This difference is due, of course, to the fact that the substitute players, as a group, are not such good batters as are the regulars.

Because it roughly resembles the letter *J*, the curve of Chart 86 is termed

a "*J* curve." Notice that the death rates from accidental causes were lower for younger persons and higher for persons who were older. The slight upturn at the very left of this curve is not always present in a *J* curve.

The shape of the curve of Chart 87 is essentially the reverse of the one just mentioned. Such a curve may be termed a "reverse *J* curve." The curve in Chart 87 indicates the length of time during which cars were parked in the Loop district of Detroit, and shows a great many cars parked
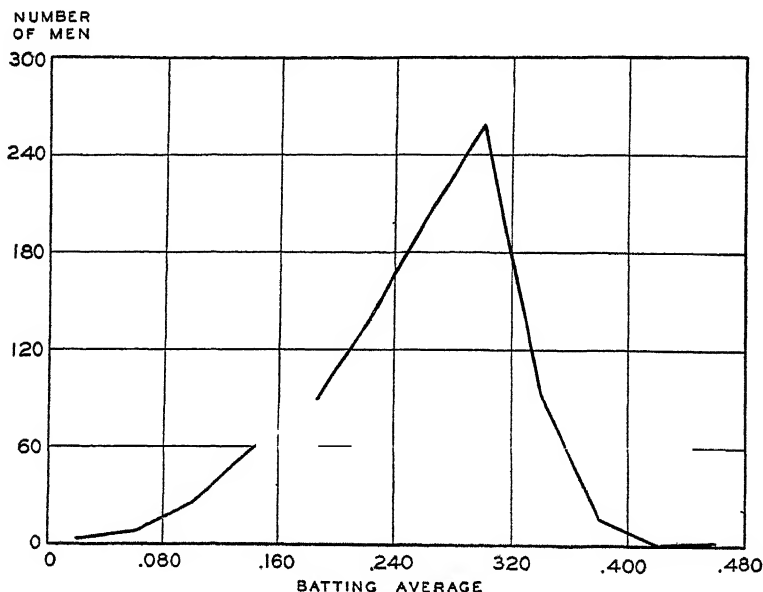
Chart 85. Batting Averages of 881 Major and Minor League Players, 1936. (Data compiled from newspaper summaries by David L. Rolbein. The players included are those who participated in 10 or more games and who were at bat 25 times or more. The minor leagues included were the International Association, the American Association, and the Pacific Coast League.)

for short periods and generally smaller numbers parked for longer lengths of time.

Another type of curve which is occasionally observed has large numbers of cases at each end of the *X*-scale and smaller frequencies for the intermediate *X* values. The term "*U* curve" is applied to frequency polygons of this shape. Chart 88 depicts a *U* curve showing for the state of Michigan the proportion of males in each age group who were unemployed in January 1935. A bimodal curve, showing two concentrations of frequencies appears as Chart 100 and is discussed in the accompanying text. This is a rather unusual form and need not be considered here.

Plotting a frequency curve when the class intervals are unequal.   In the case of certain frequency distributions it is not feasible to maintain the same class interval throughout.   The distribution of Table 29 has ten classes of 2 5 cents, eight classes of 5 cents, eight classes of 10 cents, and

RATE   PER   100,000
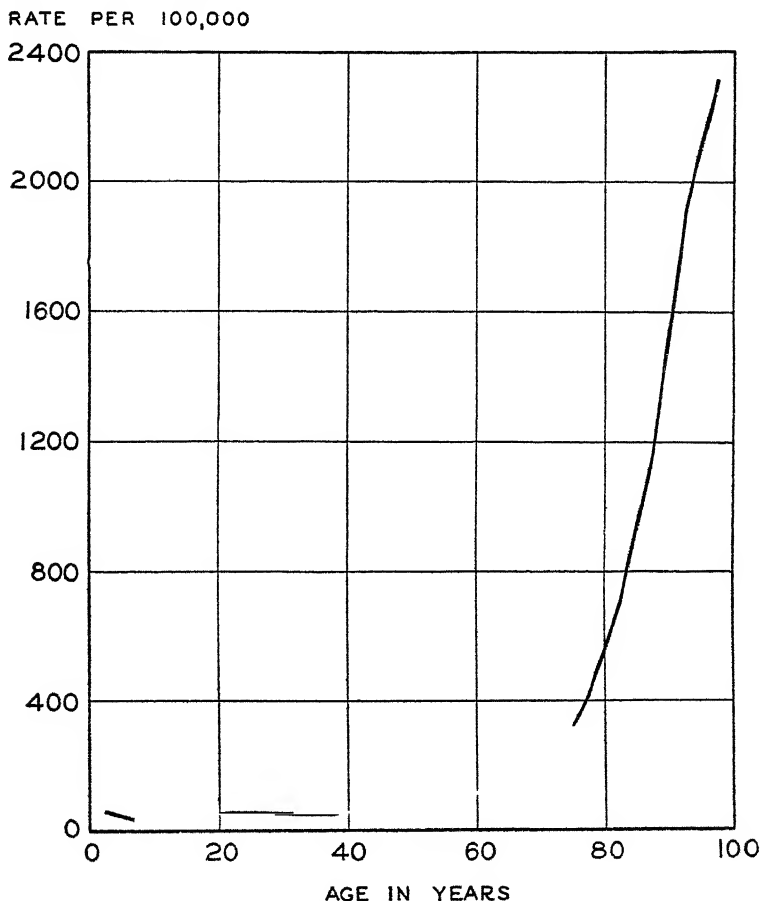
Chart 86.   Death Rates from Accidents, by 5-Year Age Groups (for the Registration States of 1920), 1933.   (Data from United States Bureau of the Census.)

one class of indeterminate width.   It would not have been desirable to use 2.5 cent intervals throughout since that would have necessitated fifty-eight classes to cover the range from 25 cents to 170 cents.   Not only would there be far too many classes to be useful, but many classes would include no, or very few, frequencies.   Class intervals of 5 cents throughout would not be desirable either, since details concerning those having average

hourly earnings of less than 50 cents would be lost, in spite of the fact that there would be twenty-nine classes instead of the original twenty-six to cover the range 25 cents to 170 cents.

When it is desired to construct a curve of data such as those in Table 29,

THOUSANDS OF CARS

HOURS PARKED
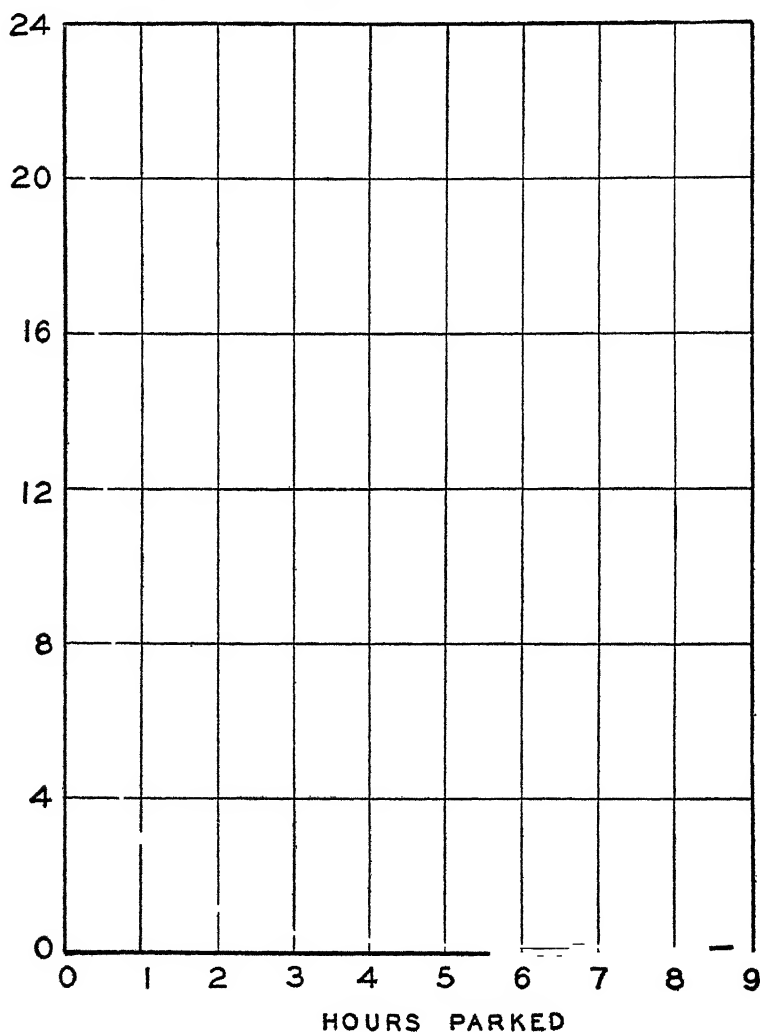
Chart 87.   Parking Time of Motor Vehicles in a Congested Area of Detroit, 1927. (Data from a chart in *Facts and Figures of the Automobile Industry,* 1928 Edition, p 84, National Automobile Chamber of Commerce, now the Automobile Manufacturers Association.)

it is necessary to make adjustments because of the varying widths of class intervals. The class "50.0 and under 55.0 cents" is twice as wide as those which precede it. We do not know how many of the 1,437 wage earners received between 50.0 and 52.5 cents per hour, and how many from 52.5 to 55.0 cents. We can say, however, that on the average there were 718.5 wage earners in each of the two halves of the class "50.0 and under 55.0

PER CENT
UNEMPLOYED
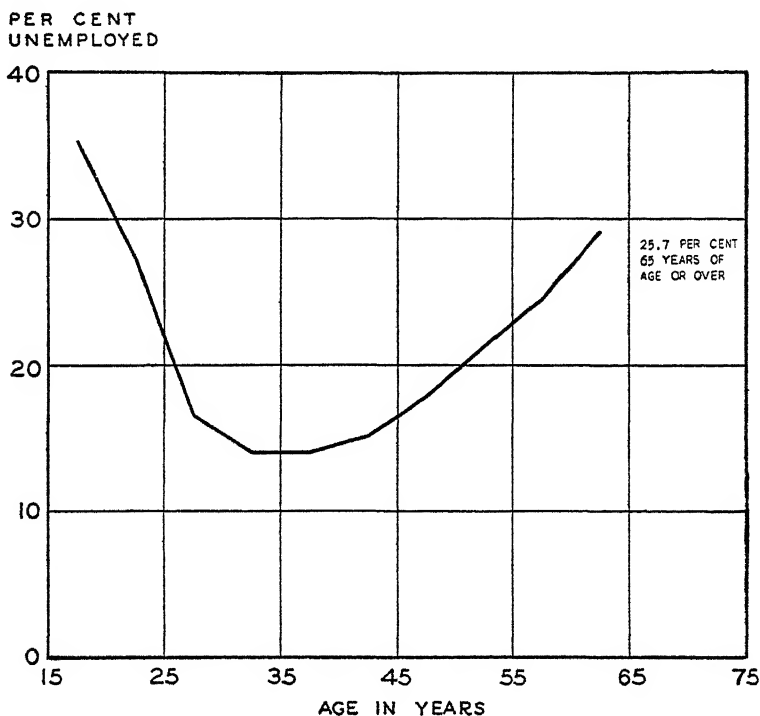
Chart 88. Per Cent of Employable Males Who Were Unemployed, by Age Groups, Michigan, January 1935. (Data from *Monthly Labor Review*, Vol. 43, No. 5, November 1936, p. 1160)

cents." Adjustments of this type have been made in the last column of Table 29 and give us frequencies for each 2.5 cents of the series. These may be thought of as frequency densities per 2.5 cents of hourly earnings.

The distribution may now be plotted in terms of the frequency densities, as in Chart 89. No estimate can well be made of the width of the last class interval of this distribution, and consequently no adjustment has been made in the table. Notice how attention was called to the presence of these 147 wage earners on the chart. An alternate method consists of putting a number or symbol at the end of the curve and showing the same

information in a footnote to the chart.   One study which necessitated the rather frequent use of open-end classes included an appendix giving further details concerning the items falling in those classes.[1]
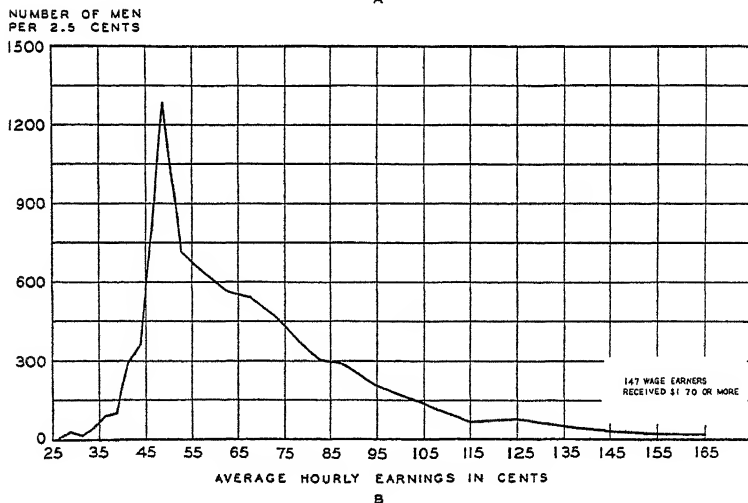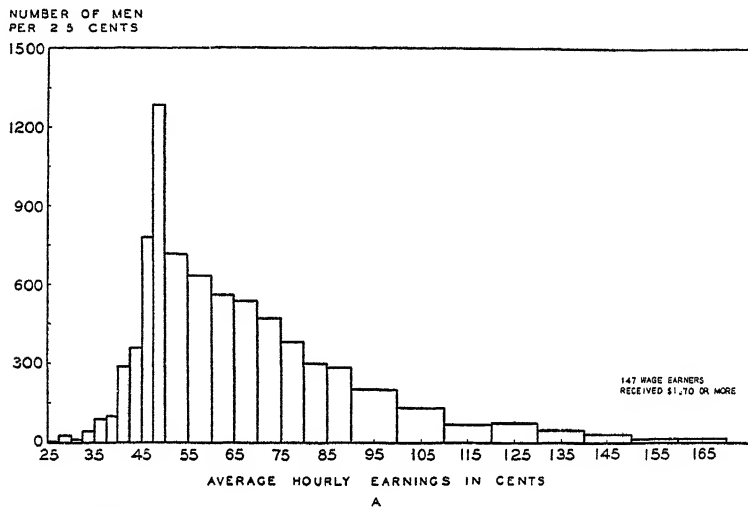
Chart 89.   Frequency Densities of Hourly Earnings of 13,427 Wage Earners in Open-Hearth Furnaces, 1935. A. Column diagram; B. Frequency curve.   (Data of Table 29)

Comparison of frequency distributions.   Table 30 shows two frequency distributions: one giving the distribution of weekly earnings of male em-

---

[1] W. A. Paton, *Corporate Profits as Shown by Audit Reports*, Appendix B (pp. 120–124), National Bureau of Economic Research, New York, 1935.

NUMBER OF EMPLOYEES

NORTH

SOUTH

162 IN THE NORTH,
1 IN THE SOUTH,
$48 AND OVER

DOLLARS PER WEEK
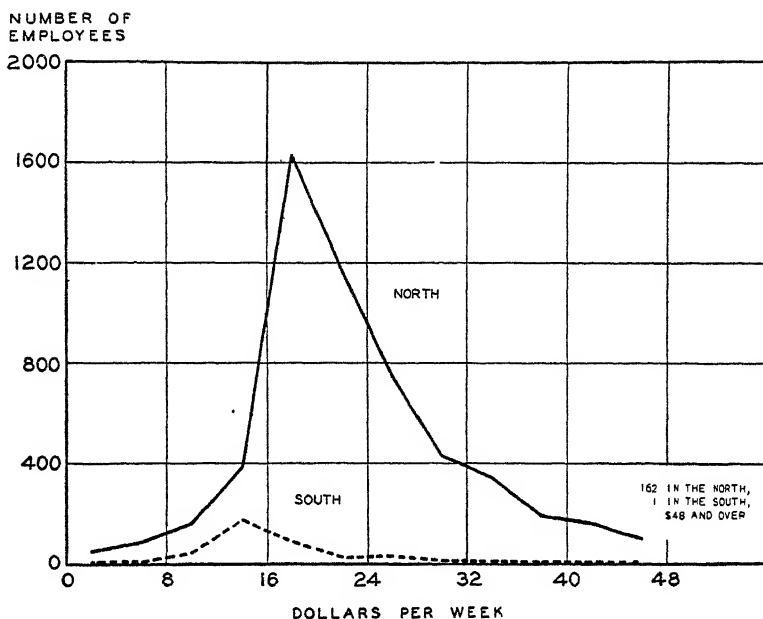
Chart 90.   Weekly Earnings of Male Employees of Folding-Paper-Box Factories in the Northern and Southern Portions of the United States, August 1935.   (Data of Table 30.)

TABLE 30

WEEKLY EARNINGS OF MALE EMPLOYEES OF FOLDING-PAPER-BOX FACTORIES IN THE NORTHERN AND SOUTHERN PORTIONS OF THE UNITED STATES, AUGUST 1935

Weekly earnings	Number of employees		Per cent of total	
	North	South	North	South
Under $ 4	49	4	.9	1 0
$ 4 but under 8	85	6	1.5	1.4
8 but under 12	160	43	2.8	10 3
12 but under 16	385	178	6.9	42 6
16 but under 20	1,628	88	29.0	21 1
20 but under 24	1,176	28	21.0	6.7
24 but under 28	742	34	13.2	8.1
28 but under 32	427	12	7.6	2.9
32 but under 36	345	11	6.1	2.6
36 but under 40	193	6	3.4	1.4
40 but under 44	163	4	2.9	1.0
44 but under 48	101	3	1.8	.7
48 and over	162	1	2.9	.2
Total	5,616	418	100.0	100.0

Source· United States Bureau of Labor Statistics, Bulletin No 620, *Wages, Hours, and Working Conditions in the Folding-Paper-Box Industry, 1933, 1934, and 1935,* pp 29, 76, and 80

ployees in the folding-paper-box industry in the northern United States
and the other presenting a distribution of weekly earnings of male em-
ployees in the same industry in the southern United States. It will be
observed that the first series refers to 5,616 males, while the second includes
but 418. If the two series dealt with approximately the same total fre-
quencies (that is, about the same number of men), we could merely plot
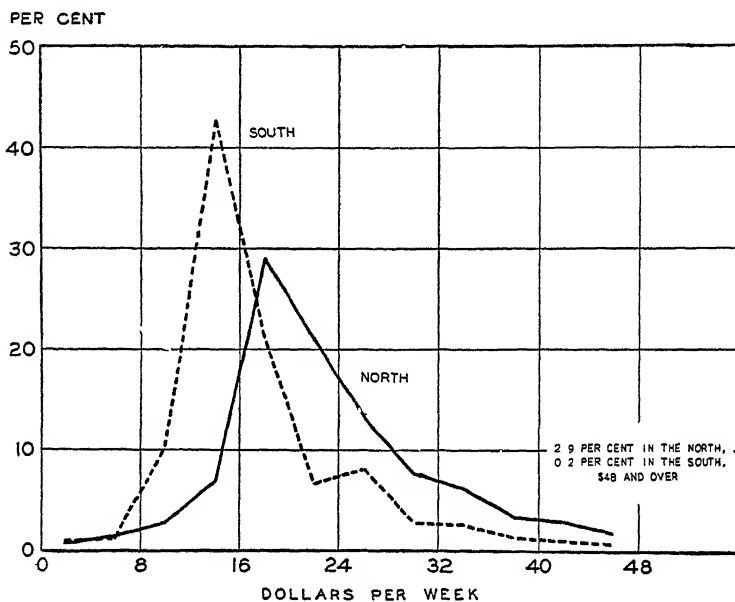
Chart 91. Percentage Distributions of Weekly Earnings of Male Employees of
Folding-Paper-Box Factories in the Northern and Southern Portions of the United
States, August 1935. (Data of Table 30.)

two frequency curves on the same axes and study their outlines. The
result of doing this for these two series is shown in Chart 90. The com-
parison is not particularly illuminating, although it is obvious that the
most prevalent earnings were between $12 and $16 per week in the South
and between $16 and $20 per week in the North. Because of the wide
difference in numbers included in the two series, we can make a more
meaningful comparison of the two curves if we express the frequencies in
each class as percentages of their respective totals. This has been done
in the last two columns of Table 30 and the resulting percentage frequency
distributions have been plotted in Chart 91. The *relative* importance of
each earnings class is now set forth clearly. Both in the $8 and under $12
and in the $12 and under $16 classes there was a larger proportion in the

South than in the North.   In each class beyond $16 there were smaller proportions in the South than in the North.

The comparison of the two series shown in Chart 91 was facilitated because the class intervals were the same in each series.   If a series of $3 class intervals is being compared with one having $6 intervals, pairs of classes of the first series may be combined, and the comparison thus made in terms of $6 intervals.   Alternately, the frequencies (or percentage fre-

NUMBER OF
MIDSHIPMEN
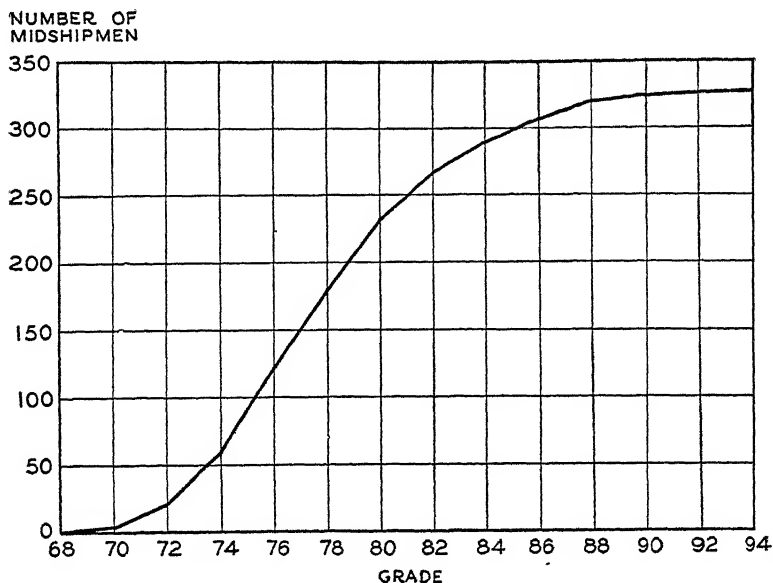
Chart 92.   Cumulative Distribution of Grades of the 1937 Graduating Class of the United States Naval Academy, Showing Number of Midshipmen Receiving Less than Stated Grade.   (Data of Table 31.)

quencies) of the series of $6 class intervals could be distributed on the basis of $3 intervals.   The frequency of each $6 class interval could be divided by 2, and this frequency assigned to each half of the interval.   It would also be possible to use two vertical scales in inverse proportion to the class intervals.   This scheme is not often used and may tend to mislead, as may the use of two (or more) vertical scales on any arithmetic chart.

Sometimes, however, the class intervals are not multiples of each other as in the instance just mentioned.   One series may have class intervals of $2, while another has intervals of $3.   We can then make the areas under the two curves the same by computing frequency densities; that is, by expressing the frequencies in each class in terms of frequencies per dollar. Thus the frequencies of each class of the first series would be divided by 2,

while the frequencies of each class of the second series would be divided
by 3. These new frequency density figures for a class interval really
refer to each dollar width of that interval, they should, of course, be plotted
at the mid-value of the class. If the number of items in the two series
is appreciably different, we may compute percentage frequencies and ex-
press these in terms of per dollar of class interval.

When two frequency distributions are expressed in terms of different
units (dollars, pounds, inches, etc.), a direct graphic comparison is not
feasible since there is no simple way in which the X-scales may be adjusted

TABLE 31

CUMULATIVE DISTRIBUTION OF GRADES OF THE 1937 GRADU-
ATING CLASS OF THE UNITED STATES NAVAL ACADEMY,
SHOWING NUMBER OF MIDSHIPMEN RECEIVING LESS
THAN STATED GRADE

Grade*	Number of midshipmen	Per cent of total
Less than 70	4	1.2
Less than 72	21	6.4
Less than 74	60	18.3
Less than 76	122	37.3
Less than 78	180	55.0
Less than 80	232	70.9
Less than 82	267	81.7
Less than 84	289	88.4
Less than 86	307	93 9
Less than 88	320	97.9
Less than 90	324	99.1
Less than 92	326	99.7
Less than 94	327	100.0

* As pointed out in Ch IX, the upper limits are 69 0 99 , 71 9499 ,
etc When rounded to whole percentage- ( ) 70, 72, etc

to each other. Certain computed values, to be discussed later, may be
used to obtain effective numerical comparison.

Cumulative frequency distributions and the ogive. The data of Table
27 show the usual (non-cumulative) form of the frequency distribution and
enable us to ascertain the number of midshipmen falling in each class.
Sometimes, however, it may be useful to know how many or what propor-
tion of students received less than certain stated grades, and this informa-
tion may be seen clearly in a cumulative table such as Table 31. In this
table the frequencies of Table 27 have been accumulated upon a "less than"
basis; we may note, for example, that 232, or 70.9 per cent, had grades
below 80, and that 60, or 18.3 per cent, had grades below 74. If the
figures of a cumulative frequency distribution are shown by means of a

curve, the result is called an *ogive*. Either the absolute or relative figures may be plotted. Chart 92 shows the ogive of the data of Table 31. When we drew a curve of the non-cumulative series, the frequencies of each class were plotted in relation to the mid-value of the class. Now, however, our cumulative frequencies in each class refer to a single value and, since 4 midshipmen had grades of less than 70, we plot the first point of the curve at 4 on the *Y*-axis and at 70 on the *X*-axis, and similarly for the other cumulative frequencies.

Instead of wishing to know how many students received less than certain specified grades, we may wish to know how many (or what proportion) received given grades or above. The data of midshipmen's grades have been cumulated upon an "or more" basis in Table 32. Now it may be

TABLE 32

Cumulative Distribution of Grades of the 1937 Graduating Class of the United States Naval Academy, Showing Number of Midshipmen Receiving Stated Grade or Above

Grade*	Number of midshipmen	Per cent of total
68 or more	327	100 0
70 or more	323	98.8
72 or more	306	93.6
74 or more	267	81.7
76 or more	205	62.7
78 or more	147	45.0
80 or more	95	29.1
82 or more	60	18.3
84 or more	38	11.6
86 or more	20	6.1
88 or more	7	2 1
90 or more	3	.9
92 or more	1	.3

* As indicated in Ch IX, the lower limits are 67 9500   , 69 9500   . ,
etc When rounded to whole percentages, these become 68, 70, etc

observed that 323, or 98.8 per cent, had grades of 70 or more; that 95, or 29.1 per cent, had grades of 80 or more; etc. The ogive for this cumulative distribution is shown in Chart 93. Again the frequencies are plotted against the stated values. For the first group, 327 on the *Y*-axis is plotted against 68 on the *X*-axis; for the second group, 323 is plotted against 70; and so on.

Cumulative frequency tables and ogives are often used to present data of wages and of hours of work. With reference to wages, they enable us to ascertain how many (or what proportion) of a group receive less than a

subsistence level, standard level, or comfort level. Similarly, we can ascertain the number or proportion receiving a subsistence level or more, a standard level or more, and a comfort level or more. It is also possible to ascertain what wage the lowest (or highest) paid 10, 25, 50, or other per cent of the workers are receiving. With respect to hours of work, we

NUMBER OF
MIDSHIPMEN
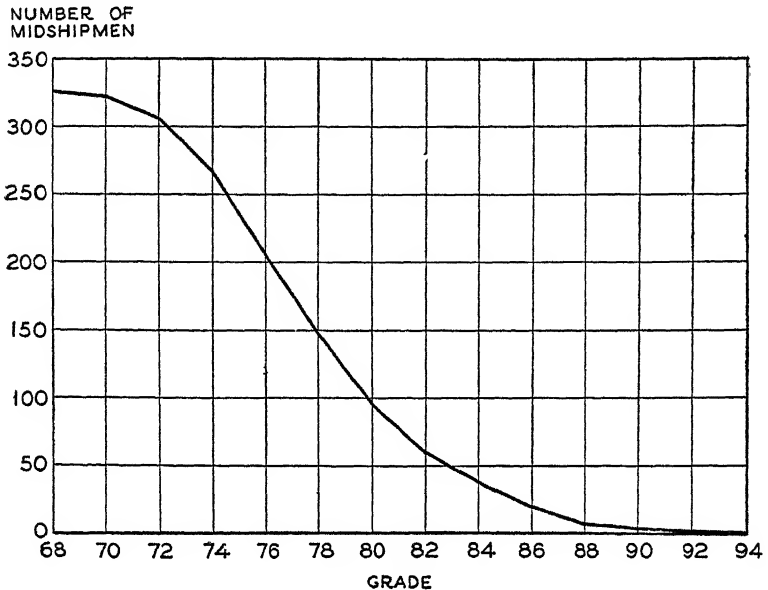
GRADE

Chart 93. Cumulative Distribution of Grades of the 1937 Graduating Class of the United States Naval Academy, Showing Number of Midshipmen Receiving Stated Grade or Above. (Data of Table 32.)

can see quickly the number or proportion working unusually long or short hours.

Comparison of ogives. If two cumulative frequency distributions are based upon nearly the same number of items, their ogives may be plotted and compared in absolute terms. If, however, the two series are based upon different totals, it is essential that comparisons be based upon the percentage frequencies. Two "or more" ogives relative to hours of work in 1933 and 1935 in the paper-box industry are shown in Chart 94. Because the 1935 ogive is steeper than that for 1933, it is apparent that a larger proportion of employees was working very long and very short hours in 1933 than in 1935, while a larger proportion was concentrated around an intermediate figure, the former NRA code level, in 1935. These conclusions would also be apparent if we examined two non-cumulative curves of the percentage frequencies, for they would differ in respect to dispersion.

TABLE 33

PRODUCTION OF ICE CREAM FOR 2,427 ESTABLISHMENTS ENGAGED PRIMARILY IN THE MANUFACTURE OF THAT PRODUCT, 1935

Production groups (gallons)	Number of establishments	Amount produced (gallons)	Per cent of		Less than cumulative per cent of	
			Number of establishments	Amount produced	Number of establishments	Amount produced
Under 5,000	144	492,084	5.9	.3	5.9	.3
5,000– 9,999	330	2,418,261	13.6	1.4	19.5	1.7
10,000– 24,999	680	11,210,317	28.0	6.4	47.5	8.1
25,000– 49,999	521	18,410,774	21.5	10.5	69 0	13 6
50,000– 99,999	370	25,744,775	15.2	14.7	84.2	33.3
100,000–249,999	268	41,497,665	11.0	23.7	95 2	57.0
250,000–499,999	70	23,392,715	2.9	13 4	98 1	70 4
500,000 and over	44	52,046,887	1.8	29 7	100.0	100.0
Total	2,427	175,213,478	100 0	100.0	.	

The ogives, however, are unique in showing cumulative comparisons such as: (1) a larger proportion of employees worked 32 hours or more in 1935; (2) about the same proportion worked 44 hours or more in the two years; (3) a larger proportion worked 48 hours or more in 1933.

The Lorenz curve. The 1935 Census of Manufactures enumerated 2,427 establishments engaged primarily in the manufacture of ice cream. The second column of figures in Table 33 shows the number of establishments grouped according to the amount of ice cream produced by the firms
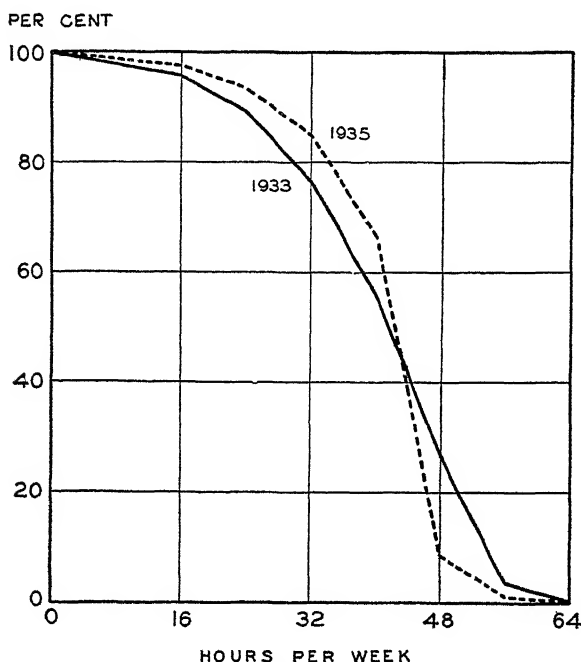
Chart 94. Cumulative Percentage Distributions of Weekly Hours of Labor of Employees of Set-Up Paper-Box Factories, 1933 and 1935, Showing Percentages Working Stated Number of Hours or More. (Based on data in United States Bureau of Labor Statistics, *Wages, Hours, and Working Conditions in the Set-Up Paper-Box Industry, 1933, 1934, and 1935,* Bulletin No. 633, p. 24.)

in each group while the third column of figures shows the amount of ice cream produced by all the firms in the group. The percentage figures are based upon these two columns. Notice that the 44 largest firms produced more ice cream than the 1,675 smallest firms comprising the first four groups. It may be seen also that the 474 small firms in the first two groups, amounting to 19.5 per cent of the total number of establishments, produced 1.7 per cent of the total output: whereas the 44 largest firms.

constituting but 1.8 per cent of all establishments, were responsible for 29.7 per cent of the ice cream produced. This tendency of production to be concentrated in a few large firms may be represented graphically by a Lorenz curve. In order to construct a Lorenz curve, the two sets

TABLE 34

ESTIMATED DISTRIBUTION OF INCOME OF 39,458,300 FAMILIES AND SINGLE INDIVIDUALS
IN THE UNITED STATES, JULY 1935–JUNE 1936

Income class		Number of families and single individuals	Number per $100 income (frequency densities)	Number receiving lower limit of each class or more
Under $	250	2,123,534	*	39,458,300
$ 250 but under	500	4,587,377	1,834,951	37,334,766
500 but under	750	5,771,960	2,308,784	32,747,389
750 but under	1,000	5,876,078	2,350,431	26,975,429
1,000 but under	1,250	4,990,995	1,996,398	21,099,351
1,250 but under	1,500	3,743,428	1,497,371	16,108,356
1,500 but under	1,750	2,889,904	1,155,962	12,364,928
1,750 but under	2,000	2,296,022	918,409	9,475,024
2,000 but under	2,250	1,704,535	681,814	7,179,002
2,250 but under	2,500	1,254,076	501,630	5,474,467
2,500 but under	3,000	1,475,474	295,095	4,220,391
3,000 but under	3,500	851,919	170,384	2,744,917
3,500 but under	4,000	502,159	100,432	1,892,998
4,000 but under	4,500	286,053	57,211	1,390,839
4,500 but under	5,000	178,138	35,628	1,104,786
5,000 but under	7,500	380,266	15,211	926,648
7,500 but under	10,000	215,642	8,626	546,382
10,000 but under	15,000	152,682	3,054	330,740
15,000 but under	20,000	67,923	1,358	178,058
20,000 but under	25,000	39,825	796 5	110,135
25,000 but under	30,000	25,583	511 7	70,310
30,000 but under	40,000	17,959	179.6	44,727
40,000 but under	50,000	8,340	83.4	26,768
50,000 but under	100,000	13,041	26.1	18,428
100,000 but under	250,000	4,144	2.76	5,387
250,000 but under	500,000	916	.366	1,243
500,000 but under	1,000,000	240	.048	327
1,000,000 and over		87	. . .	87
Total		39,458,300	. .	. . .

* No frequency density is shown for this class because the lower limit of the class is not indicated and thus the width of the class is not apparent

Source· National Resources Committee, *Consumer Incomes in the United States, their Distribution in 1935–36*, p 6

of percentages are cumulated on a "less than" basis as shown in the last two columns of Table 33. These sets of cumulative percentages are then plotted against each other as shown in Chart 95. The diagonal

line in the chart represents the line of even distribution—a condition which could not obtain unless all establishments were equally productive. The diagonal serves, however, as a basis of reference, since the more the plotted curve departs from the diagonal the less uniform is the distribution of production among the establishments of different size.

The Lorenz curve is not limited to showing the type of data in Table 33. It is also useful for showing concentration of personal or corporate income. By drawing two or more Lorenz curves, we may compare income

PER CENT OF
PRODUCTION
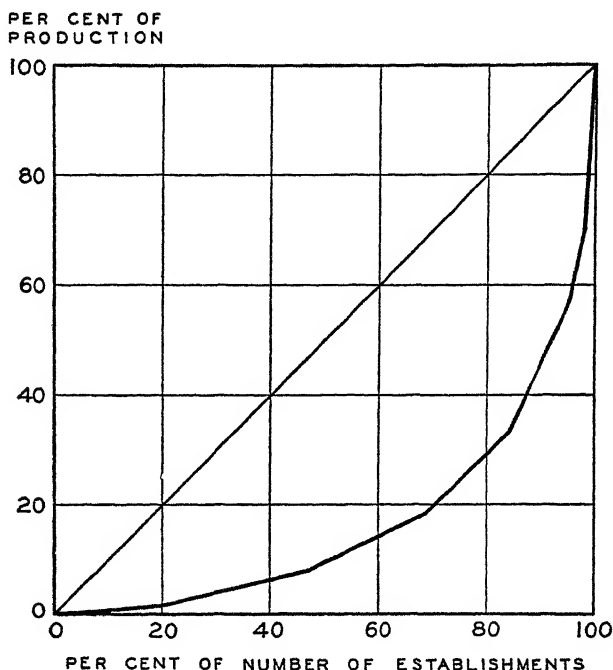
Chart 95. Lorenz Curve Showing Concentration of Production of Ice Cream, 1935.
(Data of Table 33.)

distributions at different times or places and, in the case of corporations, between different industries. To plot a Lorenz curve, we must, of course, know not only the frequencies for each class of the distribution but also the total amount for each class (as in the second and third columns of Table 33). It should be noted that class intervals of equal width are not required for the Lorenz curve.

The Pareto curve. In Table 34 are shown data of the distribution of incomes in the United States in 1935–1936 as estimated by the National Resources Committee. It will be observed, first, that the class intervals

are of varying size; second, that the number of cases falling in the various classes are very great for some classes but quite small for others; and third, that the distribution is decidedly skewed.

To draw a graph of this distribution in the conventional manner is

NUMBER
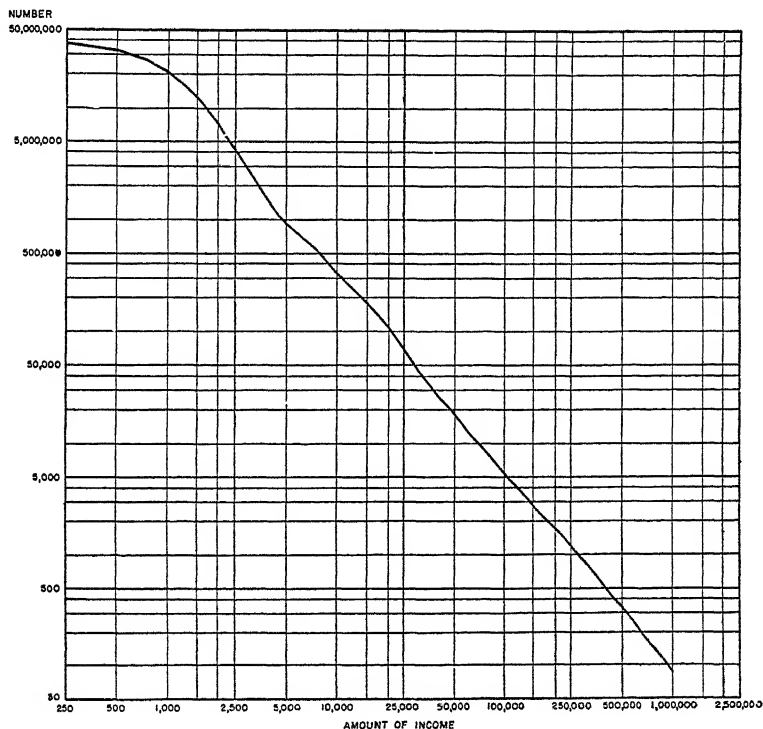
Chart 96. Pareto Curve of Cumulative Distribution of Annual Incomes of 37,334,766 Families and Single Individuals in the United States Receiving $250 or More, July 1935—June 1936. (Data of Table 34.)

virtually an impossibility. It would be necessary to use frequency densi-ties because of the varying class intervals, and these are given in the table. If the horizontal scale were 10 feet long to allow for values from 0 to $1,000,-000 only, then the horizontal distance for each of the first ten classes would be just a little less than $\frac{1}{32}$ of an inch! Furthermore, if the vertical scale were about one foot high, the curve would be about $\frac{2}{3}$ of an inch above the X-axis after passing over about $1\frac{1}{2}$ inch of the horizontal scale (at about $10,000). The reader should try to visualize this curve, as it is, of course, not feasible to include such a chart in a book of ordinary size.

One device which will enable us to depict this sort of distribution

graphically is often designated as the Pareto curve.   To draw a Pareto curve, the frequencies are first cumulated on an "or more" basis, and then an ogive is plotted on logarithmic axes.   Chart 96 shows a Pareto curve of the data of Table 34.   As in the case of all "or more" ogives, the cumulative frequencies are plotted against the lower limits of the classes.   Apparently it was because the major portion of such curves usually approximated a straight line ɯhat Pareto enunciated his law of income distribution.[2]   Although this law has been discredited as a rigid generality, the
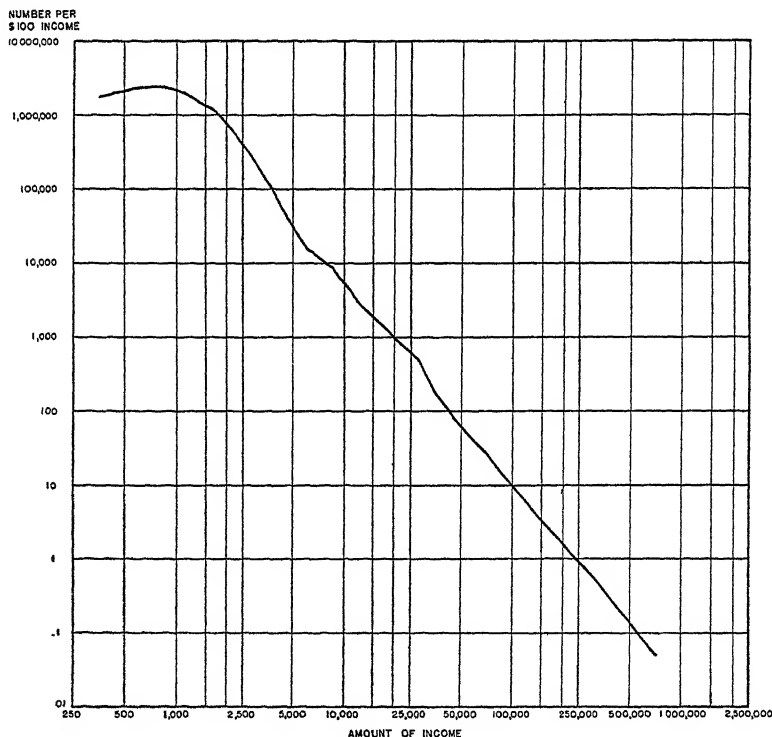
Chart 97.   Pareto Curve of Frequency Densities of Annual Incomes of 37,334,766 Families and Single Individuals in the United States, Receiving $250 to $1,000,000, July 1935—June 1936.   (Data of Table 34.)

chart is useful as a graphic device for plotting a frequency distribution which has marked positive skewness, or for comparing two or more such distributions.   Although the double logarithmic grid is most frequently used for showing income distributions, it may be used for any frequency

[2] For a discussion see *Income in the United States*, Vol. I (1921), pp. 118–124, and Vol. II (1922), Ch. 28, National Bureau of Economic Research, New York.

distribution which is so markedly skewed to the right that arithmetic plotting will not suffice.   An important characteristic of such a cumulative curve is that the steeper the curve, the more nearly equal the distribution (that is, the less the dispersion).   Two considerations in constructing such a chart as Chart 96 should be noted: first, it is not possible to show negative incomes; second, zero incomes cannot be shown.   The curve of Chart 96, therefore, begins with the number of recipients of incomes of $250 or more.

It is not necessary to cumulate the frequencies in order to plot them on the two logarithmic scales.   The non-cumulative frequency densities may be plotted to give results as shown in Chart 97.

We frequently encounter curves which are skewed to the right (although not so much as the one just considered).   The non-cumulative data may be plotted against a logarithmic X-axis and a natural Y-axis.   A series of this type is shown in Charts 125 and 126.   The curve of Chart 126 appears nearly symmetrical.   In Chapter XI a logarithmic normal curve is fitted to the series.

Selected References

R. W. Burgess: *Introduction to the Mathematics of Statistics,* Chapter IV; Houghton Mifflin Co., Boston, 1927.

R. E. Chaddock: *Principles and Methods of Statistics,* Chapter V; Houghton Mifflin Co., Boston, 1925.

F. E. Croxton and D. J. Cowden: *Practical Business Statistics,* Chapter VIII; Prentice-Hall, Inc., New York, 1934.

W. L. Crum, A. C. Patton, and A. R. Tebbutt· *Introduction to Economic Statistics,* Chapter IX; McGraw-Hill Book Co., New York, 1938.   Deals with charting frequency distributions.

E. E. Day: *Statistical Analysis,* Chapters VI, IX; Macmillan Co., New York, 1925.

F. C. Mills: *Statistical Methods Applied to Economics and Business* (Revised Edition), Chapter III; Henry Holt and Co., New York, 1938.

G. U. Yule and M. G. Kendall: *An Introduction to the Theory of Statistics* (Eleventh Edition), Chapter 6; Charles Griffin and Co., Ltd., London, 1937.

CHAPTER IX

MEASURES OF CENTRAL TENDENCY

We have seen how to construct a frequency distribution and how to draw a frequency curve. From either the classified data or the chart it is obvious that there are certain values that are frequently present and others that occur less frequently. Most of the curves that we encounter are of the type that is very roughly "bell-shaped," as shown in Charts 81, 83, 84, and 85. For such series as these charts represent, it is obvious that the more characteristic values are in the *central* part of the distributions. We therefore use the term *measures of central tendency* (or *averages*) to identify these values which may be computed in an attempt to characterize the frequency distribution. We shall discuss in this chapter the arithmetic mean, the median, the mode, and, briefly, the geometric mean and the harmonic mean.

In the following chapter we shall consider measures of dispersion, which refer to the spread of a distribution; measures of skewness, which measure the direction and amount of asymmetry; and measures of kurtosis, which indicate the degree of "peakedness" of a series.

The Arithmetic Mean

The arithmetic mean from ungrouped data. The arithmetic mean is in such constant everyday use that nearly all of us are familiar with the concept. Sometimes we refer to the arithmetic mean merely as "the average" or "the mean," but we always use the appropriate adjective when we are speaking of the geometric mean, the harmonic mean, or some other less usual mean.

The arithmetic mean of a series of items is obtained by adding the values of the items and dividing by the number of items. Suppose that in a certain small city 1-pound loaves of fresh white bread are selling for 8¢, 10¢, 11¢, and 12¢ a loaf. The arithmetic mean of these four figures would be given by

$$\frac{8¢ + 10¢ + 11¢ + 12¢}{4} = \frac{41¢}{4} = 10.25¢.$$

If we let X_1, X_2, X_3, etc., indicate the various values; N, the number of items; and $\overline{X}$, the arithmetic mean, we have

$$\overline{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_N}{N}.$$

Or, more briefly, using the summation symbol Σ, we may say

$$\overline{X} = \frac{\Sigma X}{N}.$$

The foregoing computation of the arithmetic mean involved no consideration of the fact that different amounts of bread may have been sold at the various prices. When an arithmetic mean is computed in this fashion, it may be referred to as a *simple* arithmetic mean. It is not correct to refer to this mean as an *unweighted* arithmetic mean since each of the prices was weighted equally. Let us proceed to compute a properly weighted arithmetic mean, considering the fact that there were sold 10,000 loaves at 8¢, 8,000 loaves at 10¢, 4,000 loaves at 11¢, and 1,000 loaves at 12¢. We now have

$$\overline{X} = \frac{(10,000 \times 8¢) + (8,000 \times 10¢) + (4,000 \times 11¢) + (1,000 \times 12¢)}{23,000}$$

$$= \frac{216,000¢}{23,000} = 9.39¢.$$

If we use the symbols f_1, f_2, f_3, etc., to indicate the numbers or frequencies associated with each value being averaged, we have

$$\overline{X} = \frac{f_1 X_1 + f_2 X_2 + f_3 X_3 + \cdots}{f_1 + f_2 + f_3 + \cdots} = \frac{\Sigma f X}{\Sigma f} = \frac{\Sigma f X}{N}.$$

Ordinarily an arithmetic mean is considered to be a weighted arithmetic mean, as just described, unless otherwise specified.

It should be noted that, although the arithmetic mean price of bread is 9.39¢ per loaf, no bread is actually sold at this exact price. The arithmetic mean must therefore be thought of as a computed value and not as a value which actually exists.

Properties of the arithmetic mean. One important property of the arithmetic mean is that the algebraic sum of the deviations of the various values from the mean equals zero. This is important since it will enable us to develop a method for computing $\overline{X}$ which will save an appreciable amount of time when we are dealing with a frequency distribution. Let us consider a series of five values, 6, 8, 9, 11, 14, each one of which occurs but once. Then

$$\overline{X} = \frac{6 + 8 + 9 + 11 + 14}{5} = \frac{48}{5} = 9.6.$$

Now let us compute the deviation of each value from the arithmetic mean, $x_1 = X_1 - \overline{X}$, $x_2 = X_2 - \overline{X}$, $x_3 = X_3 - \overline{X}$, etc.   We have

X	x
6	$-3\,6$
8	$-1\,6$
9	$-\ \ 6$
11	$+1\,4$
14	$+4.4$

It will be observed that $\Sigma x = 0$; this is always true for any series of values.[1]

If we compute the deviations d of the five items from some designated value which is not the arithmetic mean, the sum of these deviations Σd will not equal zero.   If the designated value is less than the arithmetic mean, there will be too many positive deviations and the sum of the deviations will be greater than zero.   If the designated value is greater than the arithmetic mean, there will be too many negative deviations and the sum of the deviations will be a negative quantity.   Since *each* of the five (N) items has been compared to a designated number which is not the true mean, the sum of the deviations will fail to equal zero by an amount which is exactly five (N) times the amount by which the designated value deviates from the actual arithmetic mean.   It is therefore possible to designate some value as an assumed mean $\overline{X}_d$, to determine the deviations from this designated value, and, by adding (algebraically) the necessary correction $\dfrac{\Sigma d}{N}$, to obtain the arithmetic mean.[2]   The process is illustrated in Table 35 where $\overline{X}_d$ is taken as 9.   Here it is observed that $\Sigma d = +3$. If we divide this figure by N, we see that $\overline{X}_d$ was too small by .6.   This is given by

$$\frac{\Sigma d}{N} = \frac{+3}{5} = +.6.$$

This is the correction to be added to the assumed mean; thus

$$\overline{X} = \overline{X}_d + \frac{\Sigma d}{N} = 9 + \frac{3}{5} = 9.6,$$

which agrees exactly with $\overline{X}$ computed by adding the values and dividing by 5.

---

[1] See Appendix B, section IX–1.   If $\Sigma x = 0$, it is obvious that $\dfrac{\Sigma x}{N} = 0$. $\dfrac{\Sigma x}{N}$ is referred to as the "first moment about the mean," or merely as the "first moment."   In the following chapter we shall have occasion to consider the second moment $\dfrac{\Sigma x^2}{N}$, the third moment $\dfrac{\Sigma x^3}{N}$, and the fourth moment $\dfrac{\Sigma x^4}{N}$.

[2] See Appendix B, section IX–2.

TABLE 35

CALCULATION OF THE ARITHMETIC MEAN, $\overline{X}$,
BY USE OF THE ASSUMED MEAN, $\overline{X}_d = 9$

X	d
6	-3
8	-1
9	0
11	$+2$
14	$+5$
	$+3$

$$\Sigma d = +3$$

$$\overline{X} = \overline{X}_d + \frac{\Sigma d}{N}$$

$$= 9 + \frac{3}{5} = 9.6.$$

In the foregoing illustration $\overline{X}_d$ was less than $\overline{X}$. Suppose we choose $\overline{X}_d$ as 13. The computations are shown in Table 36.

TABLE 36

CALCULATION OF THE ARITHMETIC MEAN, $\overline{X}$,
BY USE OF THE ASSUMED MEAN, $\overline{X}_d = 13$

X	d
6	-7
8	-5
9	-4
11	-2
14	$+1$
	-17

$$\Sigma d = -17$$

$$\overline{X} = \overline{X}_d + \frac{\Sigma d}{N}$$

$$= 13 + \frac{-17}{5} = 9.6.$$

In this case $\overline{X}_d$ was larger than $\overline{X}$, as is indicated by $\dfrac{\Sigma d}{N} = \dfrac{-17}{5} = -3.4$. The result is, as before, $\overline{X} = 13 - 3.4 = 9.6$.

A second property of the arithmetic mean, which is of importance in connection with later discussions, is that the sum of the *squared* deviations Σx^2 is *less* when the deviations are taken around $\overline{X}$ than when they are taken around any other value.

The arithmetic mean from grouped data: long method. Table 37 shows the frequency distribution of the grades of midshipmen and it is desired to ascertain the value of $\overline{X}$ for the series. When dealing with a frequency distribution, we do not ordinarily have the original data from which the frequency distribution was made. When we do have the unclassified data (as in Table 25), we can obtain the value of the arithmetic mean most accurately by totaling the values and dividing by the number of items. When we have only the frequency distribution, we must compute the mean from the grouped data. Let us proceed to compute $\overline{X}$ for the fre-

quency distribution of Table 37, and then compare our result with the arithmetic mean computed from the unclassified data.

TABLE 37

CALCULATION OF ARITHMETIC MEAN OF GROUPED DATA BY
LONG METHOD FOR GRADES OF THE 1937 GRADUATING
CLASS OF THE UNITED STATES NAVAL ACADEMY

Grade	Mid-values of classes X	Number of midshipmen f	fX
68 0–69 9	68 95	4	275 80
70 0–71 9	70 95	17	1,206 15
72 0–73 9	72 95	39	2,845 05
74.0–75 9	74 95	62	4,646 90
76 0–77 9	76 95	58	4,463 10
78 0–79 9	78 95	52	4,105.40
80 0–81 9	80 95	35	2,833 25
82 0–83 9	82 95	22	1,824.90
84 0–85 9	84 95	18	1,529 10
86 0–87 9	86.95	13	1,130.35
88 0–89 9	88 95	4	355 80
90 0–91 9	90 95	2	181.90
92 0–93.9	92 95	1	92 95
Total		327	25,490 65

$$\overline{X} = \frac{\Sigma fX}{N} = \frac{25,490.65}{327} = 77.95.$$

In computing the arithmetic mean from a frequency distribution, we take the mid-value (sometimes called the class mark) of each class as representative of that class, multiply the various mid-values by their corresponding frequencies, total these products, and divide by the total number of items. Symbolically, if X_1, X_2, $X_3 \cdots$ represent the mid-values and $f_1, f_2, f_3 \cdot \cdot$ the frequencies, then

$$\overline{X} = \frac{f_1X_1 + f_2X_2 + f_3X_3 + \cdots}{f_1 + f_2 + f_3 + \cdots} = \frac{\Sigma fX}{N}.$$

The mid-value of a class is obtained by adding the upper and lower *limits* of the class and dividing by 2. For every frequency distribution we must consider carefully what those limits are. Considering the distribution of Table 37, we might take the limits of the third class as 72.0 and 74.0, giving a mid-value of 73.0. This would be correct if the grades had each been rounded to the *last completed* tenth so that 72.0 included values ranging from exactly 72 to 72.099 $\cdots$, 72.1 included values from exactly

72.1 to 72.199 · · ·, etc., instead of having been rounded to the *nearest* tenth, as was actually done. If rounding had been to the last completed tenth, the class should have been designated "72 and under 74." Since we are dealing with a continuous variable, the *limits* of such a class would be 72 and 74, and the mid-value 73. For the midshipmen's grades, rounding was to the nearest tenth and the lowest value in the class "72.0–73 9" is 71.95, while the highest value is 73.9499 · · ·. Thus, since the variable is continuous, the class limits are 71.95 and 73.95, and the mid-value is 72.95. The mid-values have been entered in Table 37 according to this procedure.

When a class is designated (for example) "32.00–33.99," the mid-value is actually 32.995. Many statisticians would, however, state the mid-value as 33.00 since the relative discrepancy is small. In determining the mid-values for a frequency distribution it is important to know how the readings were rounded. When no information concerning the rounding is given in connection with the frequency distribution, it is probably best to assume that figures were rounded to the nearest unit given. For example, if a one-inch class is written "12.0–12.9 inches," consider the limits as 11.95 and 12.95 inches; if a five-pound class is written "10–14 pounds," consider the limits as 9.5 and 14.5 pounds. However, for discrete data, a \$2 class "\$10.00–\$11.99" has the limits \$10.00 and \$11.99, and a \$10 class "\$70–\$79" has the limits \$70 and \$79 if data were given only in whole dollars. A class should not be written "5 pounds but under 10 pounds" unless we mean exactly what we say; namely, that items in this class do not fall below 5 pounds and do not equal 10 pounds.

Considering the mid-values for the grades of midshipmen as discussed above and using the expression $\bar{X} = \dfrac{\Sigma fX}{N}$, we find that the arithmetic mean is 77.95, as shown under Table 37. From the unclassified data of Table 25, let us compute the value of $\bar{X}$ to see how nearly the figure just obtained agrees with that value. If we total all of the individual grades and divide by 327, we have

$$\bar{X} = \frac{25,486.0}{327} = 77.94.$$

The value obtained from the frequency distribution is in very close agreement with this; in fact, the error is only +.013 of one per cent. This is an unusually close agreement, but we can generally count on a difference of not more than a few per cent at most. The value of the arithmetic mean computed from a frequency distribution will generally be in close agreement with the arithmetic mean from the unclassified data if the variable is continuous and the distribution is symmetrical. If (1) the distri-

bution is skewed or if (2) the variable is discrete (or if the data are broken), or if both (1) and (2) are true, the agreement will be less close.   Likewise, close agreement cannot be expected if the data contain irregularities because an unduly small sample was used

Whatever lack of agreement is present is due to the inadequacy of the mid-value assumptions.   It is almost always true that *none of the mid-values is actually the true concentration point of its class*.   However, a glance at Chart 81 will suggest that for groups to the *left* of the group of maximum frequency the mid-value of a group is probably *less* than the mean of that group, while for groups to the *right* of the group of maximum frequency the mid-value of a group probably *exceeds* the mean of that group   Although all the mid-value assumptions are incorrect, there is a definite tendency for the errors to offset each other, provided the distribution is approximately symmetrical   Since we have the unclassified data from which the frequency distribution was made, we can compute the mean for each class and compare the class means and class mid-values.   This has been done in Table 38; it is observed that for the first 3 classes the mid-value of each class is less than the class mean, while for 7 of the last 9 classes the mid-values exceed the class means.

TABLE 38

Comparison of Class Means and Class Mid-Values for Grades of Midshipmen

Grade	Number of midshipmen	Total of grades in each class (from Table 28)	Class mean	Class mid-value
68 0–69 9	4	276 7	69 18	68 95
70.0–71 9	17	1,212 5	71 32	70 95
72.0–73 9	39	2,851 2	73.11	72 95
74 0–75 9	62	4,641 3	74 86	74 95
76.0–77.9	58	4,462 1	76 93	76 95
78 0–79.9	52	4,103 4	78 91	78 95
80.0–81.9	35	2,833 8	80 97	80 95
82 0–83 9	22	1,821 5	82 80	82 95
84 0–85.9	18	1,528 3	84 91	84 95
86 0–87.9	13	1,124 4	86.49	86 95
88.0–89.9	4	357 5	89 38	88 95
90 0–91 9	2	181.2	90 60	90.95
92.0–93 9	1	92.1	92 1	92 95

The arithmetic mean from grouped data: short methods.   In Tables 35 and 36 (see also Appendix B, section IX–2), it was shown that we could assume a value $\overline{X}_d$ for the arithmetic mean and, making use of the fact that $\Sigma x = 0$, compute the necessary correction to obtain $\overline{X}$. This method will save us appreciable time in computing the mean from a frequency dis-

tribution. The expression for $\overline{X}$ is as before, except that the symbol f is introduced because of the frequencies in the various classes. Thus

$$\overline{X} = \overline{X}_d + \frac{\Sigma fd}{N}.$$

The selected value for $\overline{X}_d$ may be the mid-value of any class. In Table 39, $\overline{X}_d$ has been taken as the mid-value of the sixth class and the computations below the table show that $\overline{X} = 77.95$, the same as found by the longer method of Table 37.

It will be observed that all of the classes of Table 39 are of the same width. When this is true, we may further shorten our computation of $\overline{X}$ by taking our deviations from $\overline{X}_d$ in *terms of class intervals d'*. Our correction $\frac{\Sigma fd'}{N}$ will then be in terms of class intervals and must be multiplied by the class interval i before being algebraically added to $\overline{X}_d$. For the mean, then,

$$\overline{X} = \overline{X}_d + \left(\frac{\Sigma fd'}{N}\right) i.$$

The computation of $\overline{X}$ by this expression is shown in Table 40 and yields the same result as given in Tables 37 and 39. This method should always

TABLE 39

CALCULATION OF ARITHMETIC MEAN OF GROUPED DATA BY SHORT METHOD FOR GRADES OF THE 1937 GRADUATING CLASS OF THE UNITED STATES NAVAL ACADEMY

(Deviations in original units)

Grade	Number of midshipmen f	Deviation from assumed mean d	fd	
68.0–69 9	4	−10	− 40	
70 0–71.9	17	− 8	−136	
72 0–73.9	39	− 6	−234	
74.0–75 9	62	− 4	−248	
76.0–77.9	58	− 2	−116	−774
78 0–79.9	52	0	0	
80 0–81.9	35	+ 2	+ 70	
82.0–83 9	22	+ 4	+ 88	
84.0–85.9	18	+ 6	+108	
86 0–87 9	13	+ 8	+104	
88.0–89.9	4	+10	+ 40	
90 0–91.9	2	+12	+ 24	
92 0 -93.9	1	+14	+ 14	+448
Total	327	. . .	. . .	−326

$$\overline{X} = \overline{X}_d + \frac{\Sigma fd}{N} = 78.95 - \frac{326}{327} = 78.95 - 1.00 = 77.95.$$

be used when a frequency distribution is made up of equal class intervals. The greater the number of classes and the greater the number of items included in a frequency distribution, the more time is saved by this procedure. The saving is especially important when the interval is some inconvenient number, such as 16.67, for example.

The arithmetic mean from grouped data having unequal class intervals. The short methods which have just been described do not greatly facilitate our computations when we are dealing with a frequency distribution having classes of varying width. For frequency distributions of this type we may use the long method, multiplying each mid-value by the corresponding frequency, summing the products, and dividing by the number of items. When classes vary in width, the distribution is invariably skewed and we must remember that, as skewness increases, the errors in our mid-value assumptions offset each other less closely. Thus the mean computed from a frequency distribution having unequal class intervals may differ markedly from the mean computed from the unclassified data. Furthermore, as will be discussed at the end of this chapter, the arithmetic mean of a decidedly skewed distribution is seldom useful.

When the arithmetic mean is computed for the frequency distribution

TABLE 40

CALCULATION OF ARITHMETIC MEAN OF GROUPED DATA BY SHORT METHOD FOR GRADES OF THE 1937 GRADUATING CLASS OF THE UNITED STATES NAVAL ACADEMY

(Deviations in class intervals)

Grade	Number of midshipmen f	Deviation from assumed mean d'	fd'	
68.0–69.9	4	−5	− 20	
70.0–71.9	17	−4	− 68	
72.0–73.9	39	−3	−117	
74.0–75.9	62	−2	−124	
76.0–77.9	58	−1	− 58	−387
78.0–79.9	52	0	0	
80.0–81.9	35	+1	+ 35	
82.0–83.9	22	+2	+ 44	
84.0–85.9	18	+3	+ 54	
86.0–87.9	13	+4	+ 52	
88.0–89.9	4	+5	+ 20	
90.0–91.9	2	+6	+ 12	
92.0–93.9	1	+7	+ 7	+224
Total	327	..	.	−163

$$\overline{X} = \overline{X}_d + \left(\frac{\Sigma fd'}{N}\right)i = 78.95 - \left(\frac{163}{327}\right)2 = 78.95 - 1.00 = 77.95$$

of Table 41, the long method gives $\overline{X}$ = \$3,009. It so happens that the unclassified data are available, and if we add the salaries paid to each of the 328 employees we obtain \$965,780, which yields an arithmetic mean of \$2,944. The mean computed from the frequency distribution is 2.2 per cent too large because of the shortcomings of the mid-value assumptions. It may be seen from Table 41, columns 2 and 6, that the mid-value

TABLE 41

CALCULATION OF ARITHMETIC MEAN OF GROUPED DATA HAVING UNEQUAL CLASS
INTERVALS AND COMPARISON OF CLASS MID-VALUES AND CLASS MEANS FOR
ANNUAL SALARIES PAID TO EMPLOYEES OF THE BOARD OF GOVERNORS
OF THE FEDERAL RESERVE SYSTEM, 1935

Salary (1)	Mid-value X^* (2)	f (3)	fX (4)	Total salaries in each class (5)	Mean salary for each class [Col 5 ÷ Col. 3] (6)
\$ 800–\$ 1,099	\$ 950	9	\$ 8,550	\$ 8,700	\$ 967
1,100– 1,399	1,250	16	20,000	20,640	1,290
1,400– 1,699	1,550	89	137,950	137,920	1,550
1,700– 1,999	1,850	41	75,850	74,070	1,807
2,000– 2,499	2,250	40	90,000	84,500	2,112
2,500– 2,999	2,750	32	88,000	85,800	2,681
3,000– 3,499	3,250	22	71,500	69,500	3,159
3,500– 3,999	3,750	15	56,250	55,300	3,687
4,000– 4,999	4,500	19	85,500	84,450	4,445
5,000– 5,999	5,500	18	99,000	95,500	5,306
6,000– 6,999	6,500	6	39,000	36,600	6,100
7,000– 7,999	7,500	4	30,000	29,800	7,450
8,000– 8,999	8,500	4	34,000	33,000	8,250
9,000– 9,999	9,500	5	47,500	45,000	9,000
10,000– 15,999	13,000	8	104,000	105,000	13,125
Total.	. . .	328	\$987,100	\$965,780	. .

* Strictly speaking, the mid-values are \$949 50, \$1,249.50, etc   We disregard this slight difference in this problem because it represents only ⅛ of one per cent of the narrowest class intervals
Source Based on data given in *Twenty-second Annual Report of the Board of Governors of the Federal Reserve System*, pp 240–243

$$\overline{X} = \frac{\$987,100}{328} = \$3,009.$$

assumption is too small for the first two groups and too large for eleven of the last twelve groups.

Sometimes a skewed distribution has an indeterminate (or open-end) group at one end, and occasionally at both ends. For example, the last class in Table 41 might have been written "\$10,000 and over." When such a class is present, there is no indication of the value which should be chosen as representative of the class. If it is assumed that the indeterminate group

has the same width as the preceding one, the mid-value will usually be too low. The use of such a mid-value may result in offsetting the upward bias of the preceding mid-values, but we can never be sure how much off-setting takes place or that it may not even overbalance the bias. The reason a class is left indeterminate is usually because it contains a few scattering items over a wide range of values.

It should be emphasized that the value of the arithmetic mean computed for a skewed distribution of unequal class intervals is only a reasonably good approximation. It becomes even less accurate when one or two indeterminate classes are present. The difficulty involved in the computation of the mean for such a distribution is completely resolved if a footnote is added to the table giving the total of the unclassified data. If this procedure is followed, a single division suffices to give the value of the arithmetic mean.

Modified forms of the arithmetic mean. Instead of computing the arithmetic mean for all of a series of items, it may occasionally suffice to make an approximation by taking the average of the smallest and largest figures. The result of such a procedure will not differ greatly from the arithmetic mean if we are dealing with a continuous variable (or a discrete variable which does not show gaps) the distribution of which is symmetrical or nearly so. For example, meteorologists have found that it is not ordinarily necessary to take hourly temperatures throughout a day and average these 24 readings to arrive at the daily mean temperature. It suffices to average only the maximum and minimum temperatures. These two readings may be obtained from the high and low points shown on the graph traced by a recording thermometer, or they may be had from a thermometer which automatically records the maximum temperature and another which automatically records the minimum temperature.

It will be recalled that the data of midshipmen's grades is skewed to the right. Consequently we should expect the average of the lowest and highest grades to exceed the arithmetic mean computed from all of the grades. Let us determine the average of these two extreme values and see how far it departs from $\overline{X}$. The highest grade shown in Table 26 is 92.1, while the lowest grade is 68.8. The average of these two grades is 80.45. The value of $\overline{X}$ computed from the unclassified data was found to be 77.94. The discrepancy resulting from averaging the extremes is 2.51, or 3.2 per cent, and indicates that we should not use this method as an approximation of $\overline{X}$ unless the distribution is symmetrical or nearly so.

A second modification of the arithmetic mean is one which will be referred to again in connection with the measurement of seasonal movements (Chapter XVII). This modification consists essentially either of ignoring certain items on the basis that they are unusual extreme values, perhaps

resulting from the introduction of a non-homogeneous or non-comparable factor into the situation, or of dropping one or more of the highest and lowest values in an array so that only the more typical values are averaged.

Suppose that a runner has competed in the 100-yard dash in ten track meets during a season and that his times were as follows.

10 2, 10.1, 10.0, 10.0, 10.1, 10.0, 9.9, 10 1, 11 4, 10.2 seconds

Now an arithmetic mean of these ten figures is 10 2 seconds, although only three races were run this slow or slower. In the race represented by the ninth figure above, the runner was spiked and limped in to finish an extremely poor last. The figure 11.4 does not indicate his running ability and could quite logically be excluded in arriving at a mean time which represents this runner's ability. If we average the other nine figures, we obtain 10.07 seconds as the arithmetic mean for this runner under normal running conditions. In like fashion, if one race had been run with a strong wind at the runner's back, his time would be abnormally short for the 100 yards and this figure too might be omitted.[3] The procedure just described differs from the one followed in measuring seasonal movements in that only the particular values for which a specific reason could be definitely assigned have been eliminated. When measuring seasonal movements, we shall drop one, two, or more items at both ends of an array in order to average the items which seem to cluster around some central value.

Averaging percentages. It was pointed out in Chapter VII that a series of percentages based on different numbers should ordinarily be averaged by weighting each percentage in proportion to its base. There are conditions, however, under which we might want to ignore the different bases and to average several percentages using a different system of weights. For example, let us assume that a student has taken two comprehensive examinations, each covering one-half of the subject matter of a course. Suppose that the first examination included 100 "true-false" questions, upon which he made 82 per cent, while the second included 150 such questions, upon which he made 88 per cent. Since each percentage represents a level of accomplishment for one-half of the work of a term, a better description of the work of the student for the term would weight the two percentages equally, resulting in an average of

$$\frac{82 + 88}{2} = 85$$

[3] A discussion of this type of modified mean when used in connection with time studies is given in F. E. Croxton and D. J. Cowden, *Practical Business Statistics*, pp. 170–176, Prentice-Hall. Inc., New York, 1934.

rather than weight the percentages according to the number of questions asked, giving

$$\frac{(100 \times 82) + (150 \times 88)}{250} = 85.6.$$

If the second examination had been based upon 10 "essay" questions, it is even more apparent that the weighting should not be determined by the number of questions included.

Averaging averages. The general outlines of the problem of averaging averages are the same as those involved in averaging percentages. If we have several averages, each referring to a category, and wish to average these averages in order to arrive at a statement compatible with that referring to the total composed of these categories, it is necessary to weight each average according to the importance of its category. For example, consider the data of Table 42, column 3, which shows the average net earnings from current operations for nine groups of banks. If we add these nine averages and divide by nine, we obtain the figure $108,895. This figure is arrived at by giving each of the nine averages the same weight, although there were varying numbers of banks in each category. The total net earnings from current operations for all of the 7,379 banks was $53,916,889, and dividing the latter figure by the former gives $7,306.83, or $7,307, as the average net earnings from current operations. If each of

TABLE 42

NET EARNINGS FROM CURRENT OPERATIONS DURING 1934 OF INSURED COMMERCIAL BANKS NOT MEMBERS OF THE FEDERAL RESERVE SYSTEM

Banks having deposits of (1)	Number of banks (2)	Average net earnings per bank from current operations (3)	Approximate total net earnings from current operations [Col 2 × Col. 3] (4)
$ 100,000 and under	1,186	$ 694	$ 823,084
100,001–$ 250,000	2,492	1,798	4,480,616
250,001– 500,000	1,720	3,563	6,128,360
500,001– 750,000	641	6,534	4,188,294
750,001– 1,000,000	380	8,727	3,316,260
1,000,001– 2,000,000	585	14,510	8,488,350
2,000,001– 5,000,000	255	33,520	8,547,600
5,000,001– 50,000,000	116	127,694	14,812,504
Over 50,000,000	4	783,011	3,132,044
Total	7,379	. .	$53,917,112

Source: *Annual Report of the Federal Deposit Insurance Corporation for the Year Ending December 31 1934,* p. 57, and by correspondence.

the nine averages in Table 42 is multiplied by the number of banks to which it refers (see column 4), if these products are totaled giving $53,-917,112, and if this result is divided by 7,379, the average is $7,306.80, or $7,307, which agrees very closely with that just given.

As in the case of percentages, there may be some instances in which the importance of each category is dependent upon some factor other than the number of items included in the category.   Suppose that 12 tires have been run on a group of test trucks unloaded except for the driver, and have shown an average mileage of 13,618 miles.   Suppose that 20 similar tires have been used on a similar group of test trucks each carrying the driver and 2,000 pounds of load, and have shown an average mileage of 12,136 miles.   The weighted average of mileage would be

$$\frac{(12 \times 13,618) + (20 \times 12,136)}{32} = 12,692 \text{ miles.}$$

What we have actually done is to assign $\frac{20}{12} = 1.67$ times as much weight to the second average as to the first.   Actually, trucks sometimes travel unloaded, sometimes loaded, sometimes partly loaded, and sometimes overloaded.   If, for the purposes of illustration, we may assume that trucks in actual use travel $\frac{1}{5}$ of their mileage unloaded and $\frac{4}{5}$ of their mileage loaded, we should arrive at our average by

$$\frac{(1 \times 13,618) + (4 \times 12,136)}{5} = 12,432 \text{ miles.}$$

It is the importance of the various load conditions in the use of the truck which should be considered in weighting rather than the number of tires tested.   If the truck travels empty $\frac{2}{3}$ of its mileage and loaded $\frac{1}{3}$ of its mileage, we should average

$$\frac{(2 \times 13,618) + (1 \times 12,136)}{3} = 13,124 \text{ miles.}$$

The Median

The median from ungrouped data.   The median is defined as that *value* which divides a distribution so that an equal number of items are on either side of it.   If we have five items, $5, $6, $7, $8, $10, it is apparent that the value of the median is $7, since there are two items below that value and two items above it.   If we have six items, 2 inches, 5 inches, 6 inches, 7 inches, 9 inches, 12 inches, it is clear that any value greater than 6 inches and less than 7 inches will satisfy our definition.   As a matter of practice, when there are an even number of items, we usually take the value of the median as halfway between the two central items.   In this instance the median would be 6.5 inches.

From what has already been said, it is obvious that the median cannot readily be located unless the data have been put into an array or, as we shall see shortly, into a frequency distribution. It will be recalled that no arranging is necessary for computing the mean, since the items of a series may be totaled no matter what their order.

The value of the median of a series may or may not coincide with the value of an existing item. When there are an odd number of items in an array, the value of the median coincides with that of one of the items; when there are an even number of items in an array, it does not.

An important property of the median, which will be referred to again, is that it is influenced by the position of the items in the array but not by the size of the items. It has already been observed that the median of $5, $6, $7, $8, $10 is $7. The two larger items may have any values greater than $7 and the two smaller items may have any values smaller than $7, yet the median remains $7.

Before proceeding to a consideration of the computation of the median for grouped data, let us compute the value of the median for the grades of the 327 midshipmen arrayed in Table 26. We want to find the value which is so located that 163 items will be on either side of it. This is, of course, the value of the 164th item,[4] and counting from either end reveals that the value of the median is 77.5. If we had an array of 200 items, we should find the value which divides the distribution so that 100 items fall below and 100 above it. This is obviously the mean of the 100th and 101st items counted from either end of the array.

The median from grouped data. To determine the value of the median of a frequency distribution, we count half of the frequencies from either end of the distribution in order to ascertain the value on either side of which half of the frequencies fall. To determine the value of the median for the grades of the midshipmen (Table 40) we first compute $\frac{N}{2} = 163.5$.

We then proceed to ascertain the value of the median. There are 122 frequencies included in the first four classes of the distribution. The estimated value of the median is therefore obtained by interpolating 41.5 frequencies (163.5 − 122) into the fifth class, assuming that the frequencies

---

[4] For ungrouped data it may seem convenient to find the value of the median by counting $\frac{N+1}{2}$ items, beginning with the highest (or lowest) item in the array. This is not the same as saying that the median is the $\left(\frac{N+1}{2}\right)$th item. Although some persons hold this concept, it is not satisfactory. The concept of the middle item as the median is unsatisfactory when the array consists of an even number of items, and must be abandoned when the median is determined from grouped data.

in that class are evenly distributed within the class.   The median, then, is given by the expression

$$\text{Med} = 75.95 + \frac{41.5}{58} \times 2 = 75.95 + 1.43 = 77.38.$$

Exactly the same result is obtained if we begin our computations from the other end of the distribution.   There are 147 frequencies included in the last 8 classes and we proceed to interpolate 16.5 frequencies (163.5 − 147) into the fifth class, *from the upper limit toward the lower limit*.   The result is

$$\text{Med} = 77.95 - \frac{16.5}{58} \times 2 = 77.95 - .57 = 77.38.$$

The value of the median is, of course, the same whether we begin our computations from one end or the other.

The value of 77.38 just obtained for the median from the frequency distribution is in very close agreement with that of 77.5 found from the array.   Unless the data contain irregularities (accounted for by the small-ness of the sample), we can expect rather close agreement when dealing with a continuous variable, and likewise for a discrete variable if the data are not broken.

We have now computed the values of the arithmetic mean and the median for the frequency distribution of midshipmen's grades.   The mean was 77.95.   The median was 77 38.   The mean exceeds the median be-cause the distribution is skewed to the right, on account of the presence of a few high grades not offset by correspondingly low grades.   If a dis-tribution is exactly symmetrical, the mean and the median are identical. If a distribution is skewed to the left, the mean will be less than the median.   This point will be treated more fully at the end of this chapter and in the following chapter.

The computation of the median from a frequency distribution of unequal class intervals does not differ from that just described.   Neither does the presence of indeterminate groups at either or both ends complicate the procedure.   Referring to Table 41, we compute the median by first de-termining $\frac{N}{2} = 164$, the number of items on either side of the median. Since there are 155 frequencies included in the first 4 classes, we must interpolate into the fifth class $\frac{9}{40}$ of the interval.   This is

$$\text{Med} = \$2,000 + \frac{9}{40} \times \$500 = \$2,000 + \$112.50 = \$2,112.50.$$

This distribution shows much skewness to the right and, as is expected, the mean of \$3,009 is much in excess of the median.   As a matter of fact, we

shall see in the following chapter that one way of measuring skewness involves consideration of the difference between the mean and the median.

If an ogive of a distribution is plotted, it is possible to obtain the value of the median graphically, as is shown in Chart 98. The process is the graphic equivalent of the computations already made and consists of the following steps: (1) Compute $\frac{N}{2}$ and locate this point on the vertical scale.

(2) Draw a perpendicular to the Y-axis at this point and extend the per-

NUMBER OF
MIDSHIPMEN

$\frac{N}{2} = 163.5$

MED = 77.4
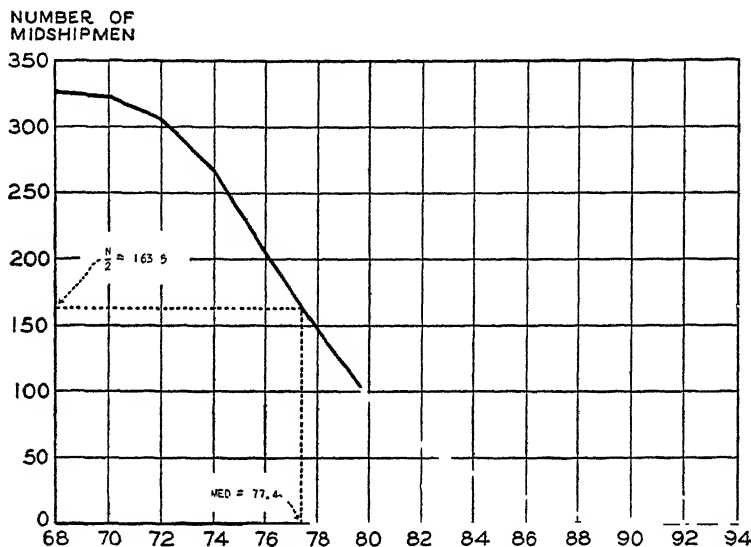
Chart 98. Graphic Location of Median for Grades of 1937 Graduating Class of the United States Naval Academy. (Data of Table 32.)

pendicular to intersect the ogive. (3) At the intersection, drop a perpendicular to the X-axis. The intersection gives the value of the median. From Chart 98 it is seen that, for the grades of midshipmen, the value of the median, located graphically, is 77.4, which is in close agreement with that computed arithmetically.

The quartiles, quintiles, deciles, and percentiles. The median characterizes a series of values because of its midway position. There are several other measures of the frequency distribution which, taken individually, are not measures of central tendency but, as we shall see later, may be used to assist in measuring dispersion and skewness. They are, however, allied to the median in that they are based upon their position in a series. We shall therefore digress at this point to discuss the *quartiles, quintiles deciles,* and *percentiles.*

There are three quartiles, which divide the distribution into four equal parts. Thus Q_1 (the first quartile, or the lower quartile) is the value located so that one-fourth of the items fall below it and three-fourths of the items exceed it. Q_2 is, of course, the median and is generally so designated. Q_3 (the third quartile, or the upper quartile) is the value so located that three-fourths of the items fall below it and one-fourth exceed it. To determine the value of Q_1 for the data of midshipmen's grades (Table 40), we count $\frac{N}{4} = \frac{327}{4} = 81.75$ frequencies from the lower limit of the first class. Thus for the value of Q_1 we have

$$73.95 + \frac{21.75}{62} \times 2 = 74.65.$$

The same result may be obtained by counting $\frac{3N}{4}$ from the upper limit of the last class.

The value of the third quartile Q_3 may be computed by counting $\frac{3N}{4}$ from the lower limit of the first class or, more expeditiously, by counting $\frac{N}{4}$ from the upper limit of the last class. Since $\frac{N}{4} = 81.75$, and since there are 60 frequencies in the last six classes, we have

$$Q_3 = 81.95 - \frac{21.75}{35} \times 2 = 80.71.$$

There are four quintiles, which divide the distribution into five equal parts; nine deciles, which divide the distribution into ten equal parts; and ninety-nine percentiles, which divide the distribution into 100 equal parts. The procedure for computing these values is similar to that for the median and the quartiles. For example, we shall compute the value of the 3rd decile, which is also the 30th percentile. We count $\frac{3N}{10} = \frac{981}{10} = 98.1$ from the lower limit of the first group and interpolate. Since there are 60 frequencies in the first 3 groups, we have

$$73.95 + \frac{38.1}{62} \times 2 = 75.18.$$

Unless a distribution is very extensive, there would be no purpose served in computing the percentiles. As a matter of fact, we generally use only the 10th, 20th, 30th, etc., percentiles, which are, of course, the 1st, 2nd, 3rd, etc., deciles.

The terms *quartile, quintile, decile,* and *percentile* are sometimes used in a different sense, to refer to the *part of the distribution* in which an item

falls. Thus, if a student is said to be in the upper quartile of his class, he is in the upper 25 per cent. If he is in the upper decile of his class, he is in the upper 10 per cent. It would undoubtedly lead to clarity of expression if we reserved quartiles, quintiles, deciles, and percentiles to mean the *measures* discussed at the opening of this section. To refer to the part of a distribution in which a student falls, we could say "highest quarter" (above Q_3), "second highest quarter" (between Q_2 and Q_3), "third highest quarter" (between Q_1 and Q_2), and "lowest quarter" (below Q_1). Similarly, we could say "fifths" in place of quintiles, "tenths" instead of deciles, and "hundredths" or "percentages" instead of percentiles.

The Mode

The mode from ungrouped data. The mode of a distribution is the value at the point around which the items tend to be most heavily concentrated. It may be regarded as the most typical of a series of values. For this very reason it is apparent that the occurrence of one or a few extremely high (or low) values has no effect upon the mode.[5] If a series of data is unclassified, not having been either arrayed or put into a frequency distribution, the mode cannot be readily located.

Taking first an extremely simple illustration: If seven men are receiving daily wages of $5, $6, $7, $7, $7, $8, $10, it is clear that the modal wage is $7 per day. If we have a series of values such as

$$3, 5, 6, 7, 9, 10, 11$$

it is apparent that there is no mode.

The mode from grouped data. If we examine the array of midshipmen's grades shown in Table 26, we find that it would be very difficult to determine the value around which the items tend to concentrate. Perhaps that value is somewhere around 76 or 77, but in coming to such a conclusion we find ourselves counting the number of grades from 75.0–75.9, from 76.0–76.9, from 77.0–77.9, and from 78.0–78.9. The mode may be located more readily by referring to a frequency distribution such as Table 40. Here it is clear that the *modal group* is 74.0–75.9; and if we take the mid-value as representative of the class, we should call 74.95 the mode.

However, there is evidence here that the mid-value is not the best esti-

---

[5] This is true in respect to the usual methods of locating the mode which are described here. If the mode is located by the expression

$$\text{Mode} = \bar{X} - \sigma\, \frac{\sqrt{\beta_1}\,(\beta_2 + 3)}{2(5\beta_2 - 6\beta_1 - 9)}$$

the extreme values do have some slight influence. The computation of the β's is discussed in the following chapter

mate of the mode. Since there are fewer frequencies in the class preceding the modal class than there are in the class following the modal class, it is logical to expect that the actual concentration is toward the upper limit of the class. We shall make use of the frequencies in these two adjacent classes to infer the probable concentration point within the modal class. The expression is

$$Mo = l_1 + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i,$$

where l_1 = the lower limit of the modal class;

   Δ_1 = the difference between the frequency of the modal class and the frequency of the preceding class (sign neglected);

   Δ_2 = the difference between the frequency of the modal class and the frequency of the following class (sign neglected);

   i = the interval of the modal class.

For the frequency distribution of grades of the midshipmen,

$$Mo = 73.95 + \frac{32 - 39}{(62 - 39) + (62 - 58)} \times 2$$

$$= 73.95 + \frac{23}{23 + 4} \times 2 = 75.65.$$

The interpolation which we have made may be illustrated graphically as shown in Chart 99. The method which we have described is sometimes called the *difference method* to distinguish it from another procedure which is more usual but less satis-
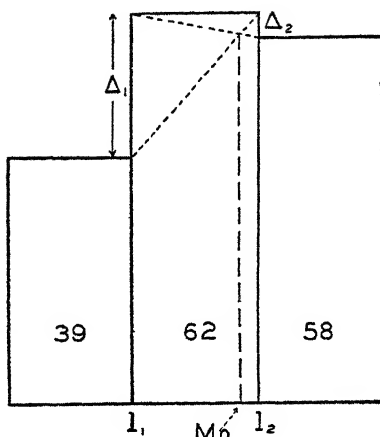
Chart 99. Diagrammatic Illustration of the "Difference Method" of Interpolating for the Modal Value. Δ_1 exerts an upward influence, and Δ_2 exerts a downward influence, each in proportion to its magnitude, so that the mode divides the interval of the modal class into two parts proportional to Δ_1 and Δ_2. That is,

$$\frac{Mo - l_1}{l_2 - Mo} = \frac{\Delta_1}{\Delta_2}.$$

Geometrically, the mode may be located by dropping a vertical line from the intersection of the two diagonals as shown on the diagram.

Algebraically the expression

$$Mo = l_1 + \frac{\Delta_1}{\Delta_1 + \Delta_2} i$$

may be developed as follows:

We wish to locate the mode so that

$$\frac{Mo - l_1}{l_2 - Mo} = \frac{\Delta_1}{\Delta_2},$$

$$\Delta_2 Mo - \Delta_2 l_1 = \Delta_1 l_2 - \Delta_1 Mo,$$
$$\Delta_1 Mo + \Delta_2 Mo = \Delta_1 l_2 + \Delta_2 l_1,$$
$$Mo(\Delta_1 + \Delta_2) = \Delta_1 l_2 + \Delta_2 l_1.$$
But $l_2 = l_1 + i.$

$$\therefore \quad Mo = \frac{\Delta_1 l_1 + \Delta_1 i + \Delta_2 l_1}{\Delta_1 + \Delta_2},$$

$$= \frac{\Delta_1 l_1 + \Delta_2 l_1}{\Delta_1 + \Delta_2} + \frac{\Delta_1 i}{\Delta_1 + \Delta_2}$$

$$= l_1 + \frac{\Delta_1}{\Delta_1 + \Delta_2} i$$

factory.[6] In any event it should be realized that we are merely making an estimate of the value of the mode. Nevertheless, it is a useful estimate and it should be remembered that the mode has two important properties; first, that it represents the most typical value of the distribution and should coincide with existing items; second, that the mode (as usually computed) is not affected by the presence of extremely large or small items.

Graphically we may obtain the mode from a column diagram as in Chart 99. We may make a very rough approximation of the mode by reading the value on the X-axis corresponding to the highest point of the curve or corresponding to the steepest portion of the ogive. The curve may be smoothed freehand since, unless the series has been subjected to a smoothing process, we should obtain a value about the same as the mid-value of the modal group.

Upon occasion, series are encountered which have two modes and are referred to as *bi-modal*. Such a series is pictured in Chart 100. Sometimes bimodality is the result of chance; sometimes it results because of the fact that two sets of non-homogeneous data are present. In Chart 100 the two concentrations are attributable to the fact that some drivers were on full (or nearly full) time work, while others were working only one or two days a week.

---

[6] The usual procedure considers the frequency of the two adjacent classes. Thus for Table 40 the 39 frequencies in the class preceding the modal class would be thought of as exerting a downward pull, while the 58 frequencies in the class following the modal class would be thought of as exerting an upward pull. Hence for the mode we should have

$$\text{Mo} = 73.95 + \frac{58}{39 + 58} \times 2 = 75.15.$$

The objection to this method is that the interpolated value for the mode is held too closely to the middle of the class. To take a rather extreme case, suppose we had a distribution showing these central classes:

Class	f
15.0–19.9	30
20.0–24 9	70
25 0–29 9	69
30.0–34.9	28

Now it is apparent that the mode is very close to 24 95 and that the difference method gives such a result:

$$19.95 + \frac{40}{40 + 1} \times 5 = 24.83.$$

The usual procedure, however, gives a figure which is clearly too low:

$$19.95 + \frac{69}{30 + 69} \times 5 = 23.43.$$

Characteristics of the Mean, Median, and Mode

Before proceeding to a consideration of other measures of central tendency, we shall examine the characteristics of these three relatively simple and very important measures.

Familiarity of the concept. The arithmetic mean is the most widely used of all the measures of central tendency. As will be pointed out later, it is frequently used under conditions which cause it to be misleading. Less well known than the arithmetic mean but very simple in concept is the idea of the median as the value which has an equal number of items on either side of it. Also less well known than the arithmetic mean, the
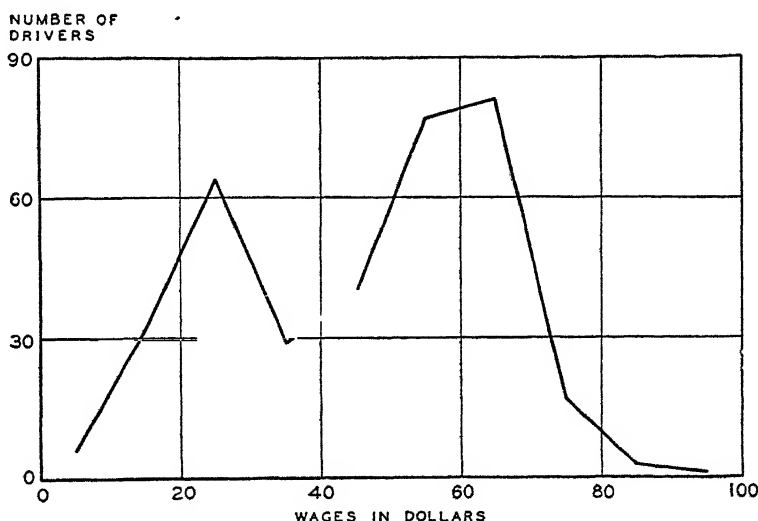
Chart 100. Distribution of Wages Received in Half Month by Drivers in Bituminous Coal Mines, Illinois, 1933. (United States Bureau of Labor Statistics, *Wages and Hours of Labor in Bituminous-Coal Mining: 1933*, Bulletin No. 601, p. 61.)

concept of the mode as the most usual or typical of a group of items is probably the simplest of the three.

The concepts of the three measures may be illustrated by means of the charts on page 216. The mean is at the point of balance, or center of gravity, such that ΣfX on one side of the mean equals ΣfX on the other side. The median divides the curve into two equal areas. The mode is the value below the peak of the curve.

Algebraic treatment. The arithmetic mean may be treated algebraically:

(a)   Since $\overline{X} = \dfrac{\Sigma X}{N}$, it follows that, if any two of the three factors (the

total, the arithmetic mean, the number of items) are known, the third may be computed. Thus

$$\overline{X} = \frac{\Sigma X}{N}; \quad \Sigma X = N\overline{X}; \quad N = \frac{\Sigma X}{\overline{X}}.$$

(b) Using appropriate weights, a series of arithmetic means may be averaged to yield the arithmetic mean of the combined distribution.

The median does not lend itself to the type of algebraic treatment discussed for the arithmetic mean. Algebraic treatment of the mode, similar to that sketched for the mean, is not possible.

The Arithmetic Mean Is at the Point of Balance or Center of Gravity.

Need for classifying data. The arithmetic mean may be computed from unclassified data, from arrayed data, from the frequency distribution, or (as noted above) merely from a knowledge of the total ΣX and the number of items N. When the arithmetic mean is computed from a frequency distribution, the value of $\overline{X}$ will very closely approximate the value of $\overline{X}$ for the unclassified data. The more nearly symmetrical the distribution, the closer the agreement of these two values.

The Areas Are Equal on Either Side of the Median.

In order that the value of the median may be computed, the data must be in an array (at least the central items must be arrayed) or in a frequency distribution. The median determined from the frequency distribution will agree approximately with that computed from the array if the distribution of items is regular within the class containing the median.
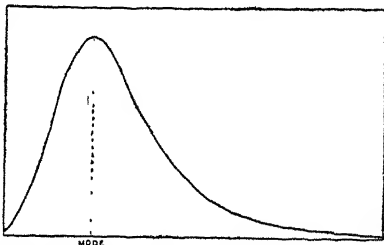
The Mode Is Below the Peak of the Curve.

The mode is most readily located from the frequency distribution, and only with some difficulty from an array. King[7] has pointed out that an array of the cities of the United States according to population of each

---

[7] Willford I. King *The Elements of Statistical Method*, p. 126, Macmillan, New York, 1919.

would show no mode. However, if such data were put into classes, a modal tendency might appear. It should be borne in mind that the process of interpolating for the modal value within the modal group is at best only an approximation. More refined methods of locating the mode involve essentially the smoothing of the data by formula and the determination of the X value of the maximum ordinate.

Effect of unequal class intervals. When classes vary in width, the value of the arithmetic mean may be computed. Such a variation of class intervals is necessitated by the presence of marked skewness (almost invariably to the right, or positive) resulting in a value for $\overline{X}$ which is not closely in agreement with that based on the unclassified data. The value of $\overline{X}$ from such a positively skewed frequency distribution would be expected to exceed the value of $\overline{X}$ from the unclassified data.

The median may ordinarily be determined rather satisfactorily from a frequency distribution having varying class intervals. The upper quartile or one or more of the upper quintiles or deciles might, however, fall in a wide class having few frequencies. The necessary interpolation would in such a case be unreliable.

When the class intervals of a frequency distribution vary in width, the mode may be satisfactorily located if the modal group and those on either side of it are of the same width. Otherwise the determination is apt to be of limited accuracy.

Effect of classes with open end. The presence of one or two indeterminate classes in a frequency distribution results in an inaccurate determination of $\overline{X}$, since mid-values ordinarily cannot be satisfactorily determined for such classes.

The presence of indeterminate classes has no effect upon the determination of the median.

Indeterminate groups do not complicate the process of locating the modal value. Occasionally, as when working with a J-, reverse J-, or U-shaped distribution, the mode is at or near the end of a distribution. Under such conditions there would not be any reason for having an indeterminate group at that end of the distribution. Incidentally, in the case of J-, reverse J-, and U-shaped distributions the mode is not a measure of *central* tendency.

Effect of skewness. For a symmetrical distribution the mean, median, and mode are identical. If the symmetrical distribution is altered on one side of the mode so as to be skewed, there is no necessary change in the value of the mode (as usually computed), but the median is changed in the direction of the skewness. Thus positive skewness (skewness to the right) increases the value of the median. The mean is increased even more. since it is affected not only by the fact that there is an excess of

frequencies on one side of the mode, but also by the amount by which the various excess frequencies deviate from the mode. Although the distribution of grades of the midshipmen is only slightly skewed, the effect of the presence of skewness is seen when we recall that the mode is 75 65, the median is 77.38, and the mean is 77.95. These values are shown on Chart 108 (p. 250).

Experience with moderately skewed distributions of continuous data indicates that the median falls about $\frac{2}{3}$ of the distance on the horizontal scale of a chart from the mode toward the mean (in the present instance the fraction is about $\frac{3}{4}$). Occasionally, therefore, the empirical formula suggested by Karl Pearson,

$$\text{Mo} = \overline{X} - 3(\overline{X} - \text{Med}),$$

is used to obtain a rough estimate of the mode. For grades of the midshipmen,

$$\text{Mo} = 77.95 - 3(77.95 - 77.38) = 76.24.$$

This estimate varies considerably from the more accurate estimate of 75.65, obtained earlier by the Δ method.

Effect of extreme values. When skewness is not general but is due to a few items deviating a great deal from the mode, the median will be only slightly affected. The arithmetic mean, however, is affected by the value of every item in the series, and the presence of one or a few extremely large (or extremely small) items in a series may result in a mean which is very misleading. As ordinarily computed, the mode is not at all influenced by the presence of a few unusually high (or low) extreme values.

The foregoing is of such great importance that we shall give further attention to it. Suppose we have the following series of seven values

$$\$12, \$14, \$15, \$15, \$16, \$18, \$19$$

the mean of which is $15.57, the median $15, and the mode $15. If an extreme value of $25 is added to these seven, the arithmetic mean becomes $16.75, the median $15.50, while the mode remains $15. Now if, instead of having added $25 as the eighth item, we add $200, the mean becomes $38.62, but the median is still $15.50 and the mode $15. The effect upon the median of any value from $16 to ∞ is the same. The mode was not at all affected by the extreme value, although, if we had added a $16 item, it would have been affected. This illustrates a different point, also; namely, that the mode is not a useful measure unless it is based upon enough items to show a well-defined concentration.

Because of the effect of extreme values upon the arithmetic mean, it is sometimes a misleading figure to use to describe a distribution. If we are considering the income of a group of people, and if most of them have

moderate incomes but one or a few have extremely high (or low) incomes, the mean will reflect these extremes and to that extent will be atypical rather than typical. An alumni association recently made a study of the graduates of 20 years ago. Among other questions asked was one concerning income during 1936. More than 350 questionnaires were sent out; only 133 replies were received concerning income of 1936. There is a large probability that these replies were selective and *any* figures derived therefrom would be of doubtful value. The mean income of the 133 replying was $13,958, but this high average was due to the fact that there were several very large incomes which were definitely extreme values. The median income was $7,500, while the mode was very close to $5,000. In such a case as this we should not use the mean alone to describe the distribution. If only one figure is to be used it is better to use the median or mode, depending upon which concept is of more importance, that of the value which has an equal number of items on either side of it or that of the most usual. It would be much better, of course, to give all three values, and, if possible, a frequency distribution or a frequency curve.

Sometimes in dealing with a series in which suspected heterogeneity is present, it may be advisable to use the median in lieu of the mean. For example, measurements might have been taken of the weight of a number of goldfish and the figures may reveal the presence of several unusually large specimens. It is suspected that, because of ignorance or carelessness, the enumerator included a few carp with the goldfish. The questionable values could be discarded. However, we are not *sure* that the heavy fish were carp, and perhaps their measurements should not be discarded. The use of the median allows the extreme values to be represented by their position in the series rather than by their size.

Sometimes we have a series in which there are present extremes of which we know the number but not the individual values. In such a situation we can determine the median or the mode, but not the mean.

When we have a series of values extending over a great range, any concept of a measure of central tendency is dubious. Suppose we have the values 4, 6, 2,000, and 2,100. It is obvious that a mean or a median could be computed but that neither would have any practical meaning.

Effect of irregularity of data. When data are broken or irregular, the value of the mean computed from a frequency distribution may be decidedly different from the value based on the unorganized data.

The same is true in the case of the median if gaps occur among the items falling in the class containing the median. When gaps occur in the vicinity of the median, the median is not a particularly good concept to use as its value would be erratic if one or two items were added to or subtracted from the series.

If a mode is clearly defined, there are not likely to be gaps near that value. When gaps are present near the mode, it is quite likely that there are too few items in the series for the mode to be either clearly defined or meaningful.

Reliability when based on samples. In Chapter XII we shall discuss the variation which may be expected in values of the mean when based on repeated random samples. At this point it will suffice to remark that the mean is more reliable than the median, and the median more reliable than the mode.

Mathematical properties. The arithmetic mean has two important properties: first, $\Sigma x = 0$; and second, $\Sigma x^2 =$ a minimum. Because of this latter property the mean is the usual basis of reference for measures of dispersion. The mean is an important function in many processes which will follow in later sections of this book. Among other uses it is essential for fitting the *normal curve* to observed data.

The sum of the deviations from the median (signs neglected) is a minimum. For this reason certain measures of dispersion are sometimes based upon the median.

Selection of appropriate measure. Using the foregoing measures as descriptive devices, the statistician may be faced with the problem of deciding which one to use to characterize a given set of data. In general, the measure of central tendency that he should use depends upon (1) the nature of the distribution of the data and (2) the concept of central tendency which is desired for a particular purpose.

If the distribution is symmetrical, or approximately so, the three measures may be used almost interchangeably. If a series is skewed, we must bear in mind that the mean is frequently not a typical value and that it may be better to use the mode (which is typical) or perhaps the median. When there are extreme deviations or when there is suspected heterogeneity, we may use the median in place of the mean, or recourse may be had to a modified mean.

If $\overline{X}$ is computed, use may be made of that value to obtain a total. Thus, if adults average 150 pounds in weight, it is safe to load about 20 people in an elevator rated to carry 3,000 pounds. (The figure of 150 pounds is somewhat high for the average weight of adults, but it is the figure frequently used to compute elevator capacity. It is obvious that the 20 people referred to should not all be heavy persons.) If subsequent computations are to be made, the mean may be required. If a curve is to be fitted to a frequency distribution, the mean will probably be used. If one series of data is eventually to be compared with another in respect to dispersion, the mean may be needed. This, however, does not mean

that the median or the mode should not be used for describing either or both of the series.

The relative standing of a person in a class may be indicated by stating whether he is better than half of the members. This rating involves the use of the median. Other statements referring to various proportions of the students may be made by using quartiles, quintiles, deciles, or percentiles.

If we are interested in knowing the typical annual expenditure of motorists for gasoline, we should make use of the mode.

Since the three measures embody different concepts, it may sometimes be advisable to use two or possibly all three. The use of the mean and the mode, or the mean and the median, gives us an idea of the amount of skewness present, as will be shown in the next chapter.

Sometimes it is necessary to make a quick estimate of the central tendency of a series. Under such conditions the mode may be promptly estimated from a frequency distribution, and the median may be quickly approximated from either an array or a frequency distribution. Of course, if the total and the number of items are given, the arithmetic mean may be computed in a few seconds.

Minor Means

The arithmetic mean, median, and mode are frequently thought of as the more important measures of central tendency, because of their wide usefulness, simplicity, and general applicability. Under certain conditions other measures of central tendency may be useful, and we shall therefore consider the geometric mean and the harmonic mean. As pointed out earlier, the term "mean" is frequently used to designate the arithmetic mean; consequently, when referring to any other mean such as the geometric mean or the harmonic mean, we should always refer to the measure by its complete designation.

The geometric mean. The geometric mean is defined as "the Nth root of the product of the items." Thus, for the four items 5, 8, 10, 12, the geometric mean is

$$G = \sqrt[4]{5 \times 8 \times 10 \times 12} = \sqrt[4]{4800} = 8.3.$$

It is interesting to note that the arithmetic mean of these four items is 8.75. For any series of positive values (not all the same), the geometric mean is smaller than the arithmetic mean.[8] When one of the values equals zero, the geometric mean equals zero and is therefore inappropriate. If

---

[8] For a demonstration, see Appendix B, section IX–3.

one or more values are negative, the geometric mean can sometimes be computed but may be meaningless. These are important drawbacks to its use.

Symbolically, the geometric mean is $\sqrt[N]{X_1 \times X_2 \times X_3 \times \cdots \times X_N}$. The computation is usually carried out by means of logarithms thus

$$\log G = \frac{\log X_1 + \log X_2 + \log X_3 + \cdots + \log X_N}{N} = \frac{\Sigma \log X}{N}.$$

The logarithm of the geometric mean is thus the arithmetic mean of the logarithms of the values.

When frequencies are present, each logarithm must be multiplied by the corresponding frequency. Thus

$$\log G = \frac{f_1 \log X_1 + f_2 \log X_2 + f_3 \log X_3 + \cdots}{N} = \frac{\Sigma f \log X}{N}.$$

For a frequency distribution, the geometric mean is usually computed by: (1) ascertaining the logarithm of the mid-value of each class, (2) multiplying each logarithmic mid-value by its proper frequency, (3) summing these products, (4) dividing by the number of items, and (5) taking the anti-logarithm of the result. The procedure is illustrated in Table 43; here it is found that $G = 77.825$, a value lower than the arithmetic mean, which is 77.95. If a series is symmetrical in a logarithmic sense (see Chapter XI) and the items are evenly distributed within the classes geometrically instead of arithmetically, it is preferable to use the mid-values of the logarithms of the class limits rather than the logarithms of the mid-values of the classes. If raw data are available, it is, of course, advisable to make the class intervals logarithmically equal also.

It will be recalled that the arithmetic mean is the sum of the values divided by the number, while the geometric mean is the Nth root of the product of the values. As noted before, N times $\overline{X}$ gives ΣX. For the geometric mean, $G^N = X_1 \cdot X_2 \cdot X_3 \cdot$ etc.; that is, the geometric mean raised to the Nth power equals the product of the values. This leads to the rather interesting point that any series of numbers having the same N and the same ΣX have the same arithmetic mean (for example, 1 and 11, 2 and 10, 4 and 8, 5 and 7, -2 and 14 all have an arithmetic mean of 6), and that any series of numbers having the same N and the same *product* have the same geometric mean (for example, 1 and 36, 2 and 18, 4 and 9 all have the geometric mean of 6).

Another property of the geometric mean is that the product of the ratios of the values on one side of the geometric mean to the geometric mean is equal to the product of the ratios of the geometric mean to the values on the other side of the geometric mean. To illustrate. let us take the values

4, 5, 20, 25, the geometric mean of which is $\sqrt[4]{10000} = 10$.   The ratios of
the values 4 and 5 to the geometric mean are $\frac{4}{10}$ and $\frac{5}{10}$, while the ratios

TABLE 43

Calculation of Geometric Mean for Grades of the 1937 Graduating Class of
the United States Naval Academy

Grade	Mid-values of classes X	$\log X$	f	$f \log X$
68.0–69.9	68 95	1 838534	4	7.354136
70 0–71 9	70 95	1 850952	17	31.466184
72 0–73 9	72 95	1 863025	39	72 657975
74 0–75 9	74 95	1 874772	62	116 235864
76 0–77 9	76 95	1 886209	58	109.400122
78 0–79.9	78 95	1 897352	52	98 662304
80 0–81 9	80 95	1 908217	35	66 787595
82 0–83 9	82 95	1 918816	22	42.213952
84 0–85 9	84 95	1 929163	18	34 724934
86 0–87 9	86.95	1 939270	13	25 210510
88.0–89.9	88 95	1 949146	4	7 796584
90.0–91 9	90 95	1 958803	2	3 917606
92.0–93 9	92.95	1 968249	1	1 968249
Total			327	618 396015

$$\log G = \frac{618\ 396015}{327} = 1.891119,$$

$$G = 77\ 825.$$

of the geometric mean to the values 20 and 25 are $\frac{10}{20}$ and $\frac{10}{25}$.   Thus we
have

$$\frac{4}{10} \cdot \frac{5}{10} = \frac{10}{20} \cdot \frac{10}{25},$$

$$\frac{1}{5} = \frac{1}{5}.$$

Similarly, we may reverse the ratios to write

$$\frac{10}{4} \cdot \frac{10}{5} = \frac{20}{10} \cdot \frac{25}{10},$$

$$5 = 5.$$

The following paragraphs discuss certain instances in which the geo-
metric mean is useful.

(1)   The geometric mean may be used for averaging ratios.   Consider the following data:

Community	Native-born inhabitants	Foreign-born inhabitants	Ratio of foreign-born to native-born (per cent)	Ratio of native-born to foreign-born (per cent)
A	8,000	4,000	50	200
B	1,500	3,000	200	50

The arithmetic mean of the two ratios of foreign-born to native-born population is 125 per cent.   Likewise, the arithmetic mean of the two ratios of native-born to foreign-born population is 125 per cent!   These two averages are inconsistent with each other.   This incongruous result does not occur if we use the geometric mean, for the geometric mean of each of the two pairs of ratios is $\sqrt{50 \cdot 200} = 100$ per cent   We could, of course, total or average the foreign-born inhabitants for the two communities, and total or average the native-born inhabitants, thus obtaining two ratios which are consistent.   There are 7,000 foreign-born and 9,500 native-born inhabitants, or an average of 3,500 foreign-born and 4,750 native-born inhabitants.   The ratio of foreign-born to native-born is

$$\frac{7,000}{9,500} \text{ or } \frac{3,500}{4,750} = 73.7 \text{ per cent,}$$

and the ratio of native-born to foreign-born is

$$\frac{9,500}{7,000} \text{ or } \frac{4,750}{3,500} = 135.7 \text{ per cent.}$$

The product of these two ratios is 1.   This arithmetic method, however, does not assign equal weight to the two ratios.   Observe that the arithmetic method involves the ratio of the means (or totals), whereas the geometric procedure involves the geometric mean of the ratios.   We have here two different concepts.   Which one to use in a given situation depends upon the purpose.   If we wish to establish a typical ratio for a number of communities and wish that ratio to be independent of the number of native-born or foreign-born persons present in the various places (that is, we wish to assign equal weight to each ratio), we may use the geometric mean of the ratios.   If we wish to allow the populations to exert an influence, we may determine the ratio of the totals or means.   The question is not whether to use an arithmetic or a geometric mean of the ratios, but whether to use a ratio based on arithmetic means (or totals) or a geometric mean of ratios.

If the two ratios of foreign-born to native-born are averaged arithmetically but weighted according to the native-born populations, the result is

73.7 per cent.   If the two ratios of native-born to foreign-born are averaged arithmetically but weighted according to the foreign-born population, we obtain 135.7 per cent.   These figures, of course, agree with those obtained by taking the ratios of the totals.

The geometric mean may be used when we wish to assign equal weight to equal ratios of change.   Suppose (a) that two commodities are selling at $2 and $10 per unit; (b) that at a later date the first commodity doubles in price while the second one is halved in price, and thus they sell for $4 and $5 respectively; and (c) that at a still later date the original price of the first commodity is halved and becomes $1, while that of the second commodity is doubled and becomes $20.   The arithmetic mean under these three situations yields: (a) $6; (b) $4 50; and (c) $10.50   The geometric mean gives: (a) $4.47; (b) $4.47; and (c) $4.47.   The assumption used to justify the geometric mean is illustrated by saying that a doubling in price offsets a halving in price, a quadrupling in price offsets a price of one-fourth the original figure, and similarly for any other two ratios whose product is 1.   This characteristic will be referred to again concerning a possible use of the geometric mean in connection with price index numbers.

(2)   Sometimes a frequency distribution is encountered which is markedly skewed to the right.   If, instead of plotting the mid-values of the classes, we use the logarithms of the mid-values (or better, plot the logarithmic mid-values, the geometric mean of each pair of limits, on a logarithmic X-scale) and a symmetrical distribution results, a geometric analysis may be proper.   This is discussed more fully in Chapter XI.

(3)   Probably the most frequently used application of the geometric principle has to do with the determination of average rates of change.   If a city had a population of 100,000 in 1920 and 120,000 in 1930, what was the average annual rate of change?   The rate of change was 20 per cent over the entire period.   If we take one-tenth[9] of that figure, or 2 per cent, as the annual rate and compute a 2 per cent increase each year over the preceding year, our 1930 population comes out as 121,900!   Obviously the correct figure is slightly smaller than 2 per cent, since we are actually compounding.   We may compute the average annual rate of change by using

$$P_n = P_o (1 + r)^n,$$

where P_o = population at beginning of period;
P_n = population at end of period;
r = rate of increase (or decrease) per year, expressed as a decimal;
n = number of years.

[9] The 1920 census was taken as of January 1, 1920, while the 1930 census was taken as of April 1, 1930.   We should, therefore, actually consider $10\frac{1}{4}$ rather than 10 years as the period between censuses.

For the data above

$$120,000 = 100,000 (1 + r)^{10}.$$

Solving this by the use of logarithms gives

$$5.079181 = 5.000000 + 10 \log (1 + r)$$

$$\log (1 + r) = \frac{.079181}{10}$$

$$= .0079181$$

$$1 + r = 1.0184$$

$$r = 1.84 \text{ per cent.}$$

The expression $P_n = P_o (1 + r)^n$ is sometimes termed the compound interest formula because of its usefulness in various problems involving compound interest. We have used it above to determine average annual rate of growth.[10] Knowing values of any three of the four symbols shown, we can solve for the fourth. Thus we may determine:

(a) Average annual rate of change r.

(b) Population a given number of years later P_n, assuming a constant rate.

(c) Number of years n until a given population will be attained, again assuming a constant rate.

(d) Population a given number of years earlier P_o, if the rate was constant.

It should be noted that the assumption of a constant rate of change for population is not valid over extended periods for any except possibly "new" countries.

The harmonic mean. The harmonic mean H is the reciprocal of the arithmetic mean of the reciprocals of the values. The expression is

$$H = \frac{1}{\dfrac{\dfrac{1}{X_1} + \dfrac{1}{X_2} + \dfrac{1}{X_3} + \cdots + \dfrac{1}{X_N}}{N}} = \frac{1}{\dfrac{\Sigma \dfrac{1}{X}}{N}}.$$

For purposes of computation, it is more convenient to use the form

$$H = \frac{N}{\dfrac{1}{X_1} + \dfrac{1}{X_2} + \dfrac{1}{X_3} + \cdots + \dfrac{1}{X_N}} = \frac{N}{\Sigma \dfrac{1}{X}}$$

---

[10] In the above discussion we found the average rate of growth between two selected points. Sometimes we wish to find the average rate of growth which best describes a number of values for different years. Such an average is not dependent upon only the first and last values of a series and is therefore more likely to be a representative figure. A method of fitting a curve to obtain such an average is given in Chapter XVI.

$$\frac{1}{H} = \frac{\dfrac{1}{X_1} + \dfrac{1}{X_2} + \dfrac{1}{X_3} + \cdots + \dfrac{1}{X_N}}{N} = \frac{\Sigma \dfrac{1}{X}}{N}.$$

The harmonic mean of the two values 3 and 12 is

$$\frac{1}{H} = \frac{\dfrac{1}{3} + \dfrac{1}{12}}{2}$$

$$= \frac{5}{24};$$

$$H = 4.8.$$

For these same values, the arithmetic mean is 7.5, while the geometric mean is $\sqrt{3 \times 12} = 6$. For any series of values (not all the same or not including zero as one value), the harmonic mean is smaller than either the geometric or the arithmetic mean.[11]

The harmonic mean is so rarely computed for a frequency distribution that we shall merely note the procedure, which consists of multiplying the reciprocal of each mid-value (or mid-value of the reciprocals of the class limits) by its frequency, adding these products, dividing by N, and taking the reciprocal of the result.

While the harmonic mean is not a measure of great importance, it is often confusing and hence we shall give a somewhat extended explanation and indicate several possible applications.

Application (1). Although oranges are not usually priced in this fashion, let us suppose that two grades of oranges are selling at 10 for \$1 and 20 for \$1. The arithmetic mean may be computed as

$$\overline{X} = \frac{10 + 20}{2} = 15.$$

That is, 15 for \$1, or \$.067 per orange. This is the price we must pay per orange *if we spend equal amounts of money for each grade.* Paying \$.067 for each of 30 oranges, we shall spend \$2.00 for the lot.

The harmonic mean gives a different result:

$$H = \frac{2}{\dfrac{1}{10} + \dfrac{1}{20}} = \frac{2}{\dfrac{3}{20}} = \frac{40}{3} = 13\tfrac{1}{3}.$$

That is, $13\tfrac{1}{3}$ for \$1, or \$.075 per orange. This is the price we must pay per orange *if equal numbers of oranges are bought at each price.* Thus, if

---

[11] See Appendix B, section IX–4.

we buy 15 oranges at 10 for \$1 and 15 oranges at 20 for \$1, we shall spend \$2.25 for all 30. Similarly, if we buy 30 oranges at \$.075 each, we shall spend \$2.25 for the lot.

The harmonic mean will give the same results as the arithmetic mean if we weight by the quantities bought at each price. Thus

$$H = \frac{30}{10\left(\frac{1}{10}\right) + 20\left(\frac{1}{20}\right)} = 15 \text{ oranges per \$1, or \$.067 per orange,}$$

assuming equal amounts of money spent for each grade.

If prices are quoted in the usual way, as so much per dozen, these oranges are selling at \$1.20 per dozen and \$.60 per dozen. A simple arithmetic mean results:

$$\overline{X} = \frac{\$1.20 + \$.60}{2} = \$.90 \text{ per dozen, or \$.075 per orange.}$$

It is the same as the first harmonic mean, since we are assuming in our computation that equal quantities are to be bought at each price. (Identical results are obtained if the quotations are per orange instead of per dozen oranges.) On the other hand, if we consider that 10 oranges may be bought at \$1.20 per dozen and 20 oranges may be bought at \$.60 per dozen, we have

$$\overline{X} = \frac{(\$1.20 \times 10) + (\$.60 \times 20)}{30} = \$.80 \text{ per dozen, or \$.067 per orange.}$$

If prices are quoted in terms of:	If the assumption is:	
	Equal amounts of money spent for each grade or commodity	Equal number of units of each grade or commodity bought at each price
Price per unit	1. $\overline{X}$, weighted by quantities for equal amounts of money (in this case units per dollar)	I $\overline{X}$, weighted by number of units (or equally)
	2 H, weighted by dollars (or equally)	II. H, weighted by dollars for equal numbers of units (or price per unit)
Units per dollar.	3. $\overline{X}$, weighted by dollars (or equally)	III $\overline{X}$, weighted by dollars for equal numbers or units (or price per unit)
	4. H, weighted by quantities for equal amounts of money (in this case units per dollar)	IV. H, weighted by number of units (or equally)

This result is the same as obtained in our first and third calculations, since we have assumed that equal amounts of money are to be spent for each grade of orange.

In the above illustrations the harmonic mean has furnished no information not already available by use of the arithmetic mean. The harmonic mean may be useful, however, when data are customarily or conveniently given in terms of problems solved per minute, miles covered per hour, units purchased per dollar, etc.

The arithmetic mean and the harmonic mean give consistent results if proper consideration is given to (a) how the data are quoted and (b) what weights are to be used. Taking prices as an illustration, the table on page 228 sets forth the relationships. Expressions 1, 2, 3, 4 give results consistent with each other. Similarly, expressions I, II, III, IV give consistent results.

Consider commodity A as selling at 4 units for $1, or $.25 each, and commodity B as selling at 10 units for $1, or $.10 each.

If equal amounts of money are to be spent for each commodity:

1. $\overline{X} = \dfrac{(.25 \times 4) + (.10 \times 10)}{14} = \dfrac{2.00}{14} = \$.1429$ per unit, or 7 for $1.

2. $H = \dfrac{2}{1\left(\dfrac{1}{.25}\right) + 1\left(\dfrac{1}{.10}\right)} = \dfrac{2}{\dfrac{7}{.50}} = \dfrac{1.00}{7} = \$.1429$ per unit, or 7 for $1.

3. $\overline{X} = \dfrac{(4 \times 1) + (10 \times 1)}{2} = \dfrac{14}{2} = 7$ for $1, or $.1429 per unit.

4. $H = \dfrac{14}{4\left(\dfrac{1}{4}\right) + 10\left(\dfrac{1}{10}\right)} = \dfrac{14}{2} = 7$ for $1, or $.1429 per unit.

If equal numbers of units of each commodity are to be bought at each price:

I. $\overline{X} = \dfrac{(.25 \times 1) + (.10 \times 1)}{2} = \dfrac{.35}{2} = \$.175$ per unit, or 5.71 for $1.

II. $H = \dfrac{.35}{.25\left(\dfrac{1}{.25}\right) + .10\left(\dfrac{1}{.10}\right)} = \dfrac{.35}{2} = \$.175$ per unit, or 5.71 for $1.

III. $\overline{X} = \dfrac{(4 \times .25) + (10 \times .10)}{.35} = \dfrac{2.00}{.35} = 5.71$ for $1, or $.175 per unit.

IV. $H = \dfrac{2}{1\left(\dfrac{1}{4}\right) + 1\left(\dfrac{1}{10}\right)} = \dfrac{2}{\dfrac{14}{40}} = \dfrac{80}{14} = 5.71$ for $1, or $.175 per unit.

From what has just been said it may be observed that (for either assumption), when averaging fractions (ratios) by the arithmetic or harmonic method, we use the arithmetic mean if weights are in the same terms as the denominator, the harmonic mean if weights are in the same terms as the numerator. Of course, if weights are in the same terms as the numerator, they may be converted into terms of the denominator and the arithmetic mean employed.

Suppose that a transaction consists of 40 handkerchiefs sold at 10 for $1 and 60 handkerchiefs sold at 20 for $1. Now we are not interested in either of the assumptions mentioned above. What we desire is the mean price when 40 handkerchiefs sell at 10 for $1 and 60 sell at 20 for $1. Using the quotations as given (that is, in terms of number of units per dollar) we may use the harmonic mean with quantity weights. Thus

$$H = \frac{100}{40\left(\frac{1}{10}\right) + 60\left(\frac{1}{20}\right)} = \frac{100}{7} = 14\frac{2}{7} \text{ per } \$1, \text{ or } \$.07 \text{ each.}$$

Still using the quotations in terms of units per dollar, we may obtain the same result by employing the arithmetic mean, if our weights are amounts of money spent for each grade. Thus

$$\bar{X} = \frac{(10 \times 4) + (20 \times 3)}{7} = \frac{100}{7} = 14\frac{2}{7} \text{ per } \$1, \text{ or } \$.07 \text{ each.}$$

If we shift our quotations to price per unit, we have 40 handkerchiefs sold at $.10 each and 60 sold at $.05 each. Now, using the harmonic mean, we weight by amounts of money spent for each grade. Thus

$$H = \frac{7}{4\left(\frac{1}{.10}\right) + 3\left(\frac{1}{.05}\right)} = \frac{7}{\frac{10}{.10}} = \$.07 \text{ each, or } 14\frac{2}{7} \text{ per } \$1.$$

Finally, using the arithmetic mean of prices per unit and weighting by quantities sold, we have

$$\bar{X} = \frac{(.10 \times 40) + (.05 \times 60)}{100} = \frac{7}{100} = \$.07 \text{ each, or } 14\frac{2}{7} \text{ per } \$1.$$

Application (2). Occasionally a frequency distribution may be encountered which is so skewed to the right that, when plotted in terms of the reciprocals of the class mid-values, it assumes an approximately normal form. In such instances harmonic treatment may be indicated. Such cases are rather unusual, however, and will not be treated in this book.

Application (3).   An interesting and apparently valid application of the harmonic mean is given in an article by Holbrook Working [12]   In his study of the factors influencing the price of potatoes, Working uses the harmonic mean, because, as he points out, a low price during part of a season will be compensated only by a disproportionally high price during the remainder of the season.   To illustrate we have selected the monthly prices for one crop year and have shown them in Chart 101.   When the reciprocals or the logarithms are plotted, the curve is straighter than when the arithmetic values are plotted, the reciprocals giving perhaps the most nearly straight line. This indicates that the harmonic mean is not inappropriate as a measure of the average price of potatoes during a season.[13]

It is sometimes argued that the geometric mean should be used for series of data having a definite lower limit and an indefinite upper limit. One type of such data is price relatives, which, having a base of 100, may fall to 0 but rise to ∞.   The question is not so much one of the existence of such limits as it is one of what values may actually occur and how they are approached—arithmetically, geometrically, or reciprocally— whether, if we are dealing with a frequency distribution, the series is approximately symmetrical in terms of X, skewed but approximately symmetrical in terms of $\log X$, or skewed but approximately normal in terms of $\frac{1}{X}$.
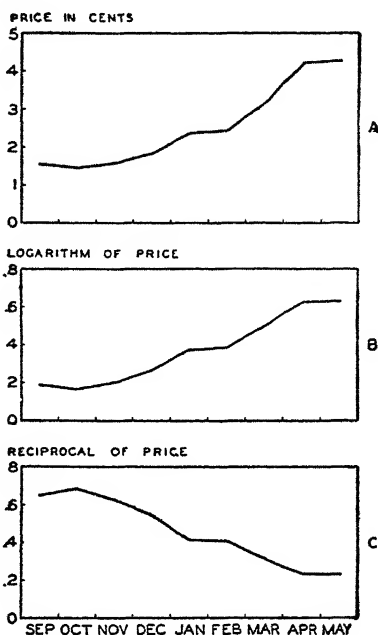
Chart 101.   Price of Potatoes per Bushel in Minneapolis and St. Paul, September 1919–May 1920:   A. Price, B. Logarithm of Price, C. Reciprocal of Price. (Data from Holbrook Working, *ibid.*, p. 40.)

[12] Holbrook Working, *Factors Determining the Price of Potatoes in St. Paul and Minneapolis,* Technical Bulletin 10, University of Minnesota Agricultural Experiment Station, pp 9 and 10.

[13] For a further discussion of the harmonic mean, see R. W. Burgess, *Introduction to the Mathematics of Statistics,* pp. 90–96, Houghton Mifflin Co., Boston, 1927.

In an arithmetic sense, a price drop of 33.3 per cent is offset by a price rise of 33.3 per cent, a decline of 50 per cent is offset by a rise of 50 per cent, and a fall of 90 per cent is offset by a rise of 90 per cent. Thus

$$\frac{66.7 + 133.3}{2} = 100,$$

$$\frac{50 + 150}{2} = 100,$$

$$\frac{10 + 190}{2} = 100.$$

In a geometric sense, a price drop of 33.3 per cent is offset by a rise of 50 per cent, a fall of 50 per cent is offset by a rise of 100 per cent, and a drop of 90 per cent is offset by a rise of 900 per cent. Thus

$$\sqrt{66.7 \times 150} = 100,$$
$$\sqrt{50 \times 200} = 100,$$
$$\sqrt{10 \times 1000} = 100.$$

In a reciprocal sense, a price drop of 33.3 per cent is offset by a rise of 100 per cent, a fall of 50 per cent is offset by a rise to ∞, and a fall of more than 50 per cent cannot be offset by any rise however great. Thus

$$\frac{2}{\dfrac{1}{66.7} + \dfrac{1}{200}} = 100,$$

$$\frac{2}{\dfrac{1}{50} + \dfrac{1}{\infty}} = 100.$$

There are a number of other measures of central tendency which are of mathematical and theoretical rather than of practical interest. The quadratic mean

$$\sqrt{\frac{\Sigma X^2}{N}}$$

is the square root of the arithmetic mean of the squares of the values Unless all the values are the same, the quadratic mean exceeds the arithmetic mean. The quadratic mean is mentioned here because the *concept* is important. Although we do not use the term "quadratic" or "mean," we shall shortly compute the quadratic mean of the *deviations* from the arithmetic mean. It will not be a measure of central tendency, but a measure of dispersion; we shall call it the standard deviation, or σ, and its expression is

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}}.$$

Selected References

R. W. Burgess: *Introduction to the Mathematics of Statistics,* Chapter V; Houghton Mifflin Co., Boston, 1927.   Includes a discussion of the harmonic mean.

R. E. Chaddock: *Principles and Methods of Statistics,* Chapters VI–VIII; Houghton Mifflin Co., Boston, 1925.

F. E. Croxton and D. J. Cowden: *Practical Business Statistics,* Chapter IX; Prentice-Hall, Inc., New York, 1934.

W. L. Crum, A. C. Patton, and A. R. Tebbutt· *Introduction to Economic Statistics,* Chapters X, XI; McGraw-Hill Book Co., New York, 1938.

F. C. Mills. *Statistical Methods Applied to Economics and Business* (Revised Edition), Chapter IV, Henry Holt and Co., New York, 1938   The characteristics of the various measures are discussed on pages 133-136

C. M Walsh. *The Problem of Estimation, A Seventeenth Century Controversy and Its Bearing on Modern Statistical Problems, Especially Index Numbers,* P. S. King and Son, London, 1921.   Averages are discussed

A. E. Waugh: *Elements of Statistical Method,* Chapter IV; McGraw-Hill Book Co., New York, 1938.

G. U. Yule and M. G. Kendall· *An Introduction to the Theory of Statistics* (Eleventh Edition), Chapter 7; Charles Griffin and Co , Ltd., London, 1937.

F. Zizek *Statistical Averages, A Methodological Study;* Henry Holt and Co., New York, 1913.

CHAPTER X

DISPERSION, SKEWNESS, AND KURTOSIS

In the preceding chapter we considered certain measures which attempted to describe the central tendency of a frequency distribution. There are other aspects of frequency distributions which are also important. First we shall consider the *dispersion*, or spread of the data. Two counties may each show an average yield of wheat of 15 bushels to the acre; but, if the data are considered farm by farm, one county may exhibit extreme values ranging from 10 to 20 bushels per acre, while the other may show yields as low as 5 bushels per acre and as high as 25 bushels per acre. If such a crude measure of dispersion may be used, it is apparent that there is greater uniformity of yield in the first county. Chart 102 shows two symmetrical curves which have the same mean but which differ in respect to dispersion.

Chart 102. Two Frequency Curves Having Different Dispersions.

If a frequency curve or frequency distribution is not symmetrical, it is said to be *skewed*, or *asymmetrical*. Most frequency distributions exhibit
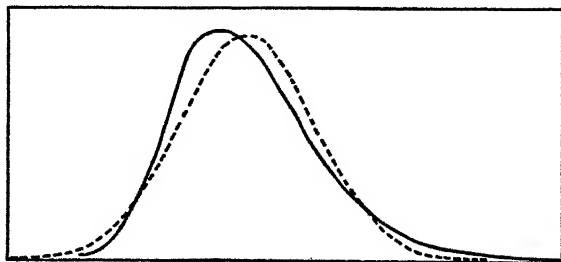
Chart 103. A Curve Skewed to the Right (Solid Line) and a Symmetrical Curve (Dashed Line).

more or less skewness.   Chart 103 shows two curves, one of which is symmetrical and one of which is skewed.   The skewed curve is skewed to the right—the direction in which the excess tail appears.

A measure of *kurtosis* indicates the degree to which a curve of a frequency distribution is peaked or flat-topped.   Our basis of reference is the normal curve described and discussed in Chapter XI.   Chart 104 shows a normal curve and a curve which is more peaked than normal; Chart 105 shows a normal curve and a flat-topped curve.

Measures of Absolute Dispersion

The mean annual temperature at Boise, Idaho, is 50.9 degrees.   The mean annual temperature at Seattle, Washington, is 51.0 degrees or almost exactly the same as at Boise.   These two figures do not, however, suffice to characterize this aspect of the climatic

Chart 104.   A Peaked or Leptokurtic Curve (Solid Line) and a Normal or Mesokurtic Curve (Dashed Line).

conditions of the two cities.   The temperature at Boise has been known to fall as low as −28 degrees and to rise as high as 121 degrees.   In Seattle the lowest recorded temperature is 3 degrees and the highest is 98 degrees.   It is quite apparent that there is greater variability of temperature at Boise than at Seattle.
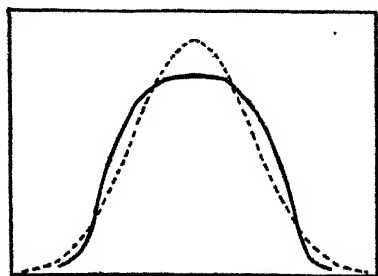
Chart 105.   A Flat-Topped or Platykurtic Curve (Solid Line) and a Normal or Mesokurtic Curve (Dashed Line).

Let us consider a second illustration. A buyer for a large department store has been offered two types of electric lights for use in the store.   The salesmen each claim about the same average length of life for their bulbs.   The buyer obtains from a testing laboratory test data for 40-watt lamps of the two makes and finds that the average life of each of the two kinds of bulbs is about 1,000 hours.   Examining the data further, however, shows that in one batch of bulbs a lamp burned out at 325 hours while one lasted 1,570 hours.   In the other batch one lamp lasted but 105 hours, while one did not burn out until the expiration of 2,910 hours.   This limited information indicates a greater degree of uniformity among lamps of the first batch

The range. The measurement of dispersion may be made in a crude form by referring to the lowest and the highest values, as was done in the preceding paragraphs. This is a very simple and easy-to-understand measure. The range gives a comprehensive value for the data in that it includes the limits within which all of the items occurred. However, the range has one important disadvantage. Because the range is based upon the two extreme items of a series, it is misleading if either one (or both) of those extremes is an unusual occurrence. The range is occasionally used for continuous data when we have a large number of items, but should be avoided when we are handling broken data. Another objection to the use of the range as a measure of dispersion arises when we are comparing two distributions with different total frequencies. The series which

TABLE 44

CALCULATION OF THE AVERAGE DEVIATION FOR GRADES OF THE 1937 GRADUATING CLASS OF THE UNITED STATES NAVAL ACADEMY

($\overline{X} = 77.95$)

Grade	Mid-values of classes X	Number of midshipmen f	Deviation of mid-values from $\overline{X}$ $\|x\| = X - \overline{X}$ (signs neglected)	$f\|x\|$
68 0–69 9	68 95	4	9 00	36
70 0–71 9	70 95	17	7 00	119
72 0–73 9	72.95	39	5.00	195
74 0–75.9	74 95	62	3 00	186
76 0–77.9	76.95	58	1 00	58
78 0–79 9	78.95	52	1 00	52
80 0–81 9	80.95	35	3 00	105
82.0–83 9	82.95	22	5.00	110
84 0–85 9	84 95	18	7 00	126
86 0–87 9	86.95	13	9.00	117
88 0–89.9	88.95	4	11 00	44
90 0–91 9	90.95	2	13 00	26
92 0–93.9	92 95	1	15 00	15
Total		327		1,189

$$\text{AD} = \frac{\Sigma f\|x\|}{N} = \frac{1,189}{327} = 3\,64.$$

includes the larger number of items is more likely to include some very low or high values or both; therefore the range is apt to increase as N increases.

Referring to the midshipmen's grades in Table 44, it is observed that the range is 67.95 (the lower limit of the first class) to 93.95 (the upper limit of the last class). If we have the array to refer to as in Table 26,

the range may be given a little more accurately as 68.75 to 92.15.   The range from the frequency distribution merely tells us that no one in the 1937 class received a grade below 67.95 or above 93.95.   The range is usually stated as the difference between the two extreme values.   For the midshipmen, $93.95 - 67.95 = 26.00$.   However, if only this single figure is given, we do not know whether the range is from 0 to 26, or from 70 to 96, or what the limits may be.

The 10–90 percentile range.   Sometimes we are interested in knowing the range within which a certain proportion of the items fall.   One such range, which is frequently used in educational measurements, is the 10–90 percentile range.   This measure excludes the lowest 10 per cent and the highest 10 per cent, giving the two values between which the central 80 per cent of the items occur.   Of course, the 10th percentile is the 1st decile, and the 90th percentile is the 9th decile.   The measure is usually referred to, however, as the 10–90 percentile range rather than the 1–9 decile range, since the former carries more clearly the idea of the central 80 per cent.

For the midshipmen's grades, we interpolate $\dfrac{N}{10}$ from the lower limit of the series to obtain P_{10}:

$$P_{10} = 71.95 + \frac{11.7}{39} \times 2 = 72.55;$$

and $\dfrac{N}{10}$ from the upper limit of the series to obtain P_{90}:

$$P_{90} = 85.95 - \frac{12.7}{18} \times 2 = 84.54.$$

The 10–90 percentile range is thus 72.55 to 84.54, or 11.99; and we know that 10 per cent had grades below 72.55, 10 per cent had grades above 84.54, while 80 per cent had grades between these values.

The quartile deviation.   In Chapter IX mention was made of Q_1 and Q_3, the lower and the upper quartiles.   A measure of dispersion based upon these values is termed the *quartile deviation*, or the *semi-interquartile range*.   It is given by

$$Q = \frac{Q_3 - Q_1}{2}.$$

If a series is symmetrical, it is clear that Q_1 and Q_3 are equidistant from the median.   Therefore, if we measure $\pm Q$ from the median, we include 50 per cent of the items of the series, for we have measured back to Q_1 and Q_3.   If a series is skewed, as is usually true, we may take $\pm Q$ around the median. and, while we shall not arrive at either Q_1 or Q_3, we may

expect to include approximately 50 per cent of the items unless the skewness is great.

In the preceding chapter it was found, for the midshipmen's grades, that
$$Q_1 = 74.65; \text{ Med} = 77.38; Q_3 = 80.71.$$
Computing the quartile deviation gives
$$Q = \frac{80.71 - 74\ 65}{2} = 3.03.$$

We therefore expect to find about half of the midshipmen's grades within $77.38 \doteq 3.03$, or between 74.35 and 80.41. Let us interpolate into the proper classes of Table 44 to see if this is true. First we want to know about how many of the 62 cadets in the fourth class had grades from 74.35 to 75.95. That is given by
$$\frac{75.95 - 74.35}{2.00} \times 62 = \frac{1.60}{2.00} \times 62 = 49.6.$$

We include all cadets in the next two classes $58 + 52 = 110$, and for the seventh class
$$\frac{.46}{2.00} \times 35 = 8.$$

The number included between 74.35 and 80.41 is therefore
$$49.6 + 110 + 8 = 167.6, \text{ or } 51.3 \text{ per cent.}$$

Neither the 10–90 percentile range nor the quartile deviation is affected by extreme values as is the range. However, these two measures have another shortcoming, in that they do not consider all of the values in measuring dispersion. The values below Q_1 (or above Q_3) could be massed closely together or spread out widely; the effect upon Q would be the same.

Occasionally, use is made of the inter-quartile range, Q_3-Q_1. For the midshipmen's grades,
$$Q_3 - Q_1 = 80.71 - 74.65 = 6.06.$$

The average deviation. The *average deviation,* or the *mean deviation* as it is sometimes called, is usually measured in relation to the arithmetic mean. The average deviation is obtained by taking the sum of the deviations of the items from the arithmetic mean, without regard to signs, and dividing by the number of items. It will be recalled that $\Sigma x = 0$, and it is for this reason that the signs of the various x values are neglected. Thus
$$AD = \frac{\Sigma |x|}{N},$$
or for a frequency distribution
$$AD = \frac{\Sigma f|x|}{N},$$

where | | means that the signs are neglected. Table 44 indicates the computation[1] of AD for grades of the midshipmen, and AD is found to be 3.64. Observe that this value is a little larger than Q, which was found to be 3.03.

If a distribution is normal, 57.5 per cent of the items are included within the range of $\overline{X} \pm AD$. If the distribution is moderately skewed, this will be found to be approximately true. For the midshipmen's grades,

$$\overline{X} \pm AD = 77\ 95 \pm 3.64 = 74.31 \text{ and } 81.59,$$

which is a slightly wider range of values than was found for the median $\pm Q$.

Let us interpolate into the frequency distribution to see what percentage of the midshipmen fall within the limits of 74.31 and 81.59. The procedure is similar to that employed in the preceding section on Q. First we estimate the number of frequencies from 74.31 to 75.95 in the fourth class of Table 44. Thus

$$\frac{1.64}{2.00} \times 62 = 51.$$

Our range of values includes all of the frequencies of the fifth and sixth classes, $58 + 52 = 110$. Finally, we estimate the number of frequencies from 79.95 to 81.59 in the seventh class. This is

$$\frac{1.64}{2.00} \times 35 = 29.$$

Combining, $51 + 110 + 29 = 190$, or 58.1 per cent of all the midshipmen. This distribution, it will be remembered, is slightly skewed (see Chart 81 or 108). The reader may wonder why we do not refer to the array of Table 26 rather than to the frequency distribution to ascertain the number of grades between 74 31 and 81.59. We could do that, but to be consistent we should also have computed the mean deviation from the data of Table 26 or 25. To do this, we use the mean of the unclassified data (77.94) and the deviations of each grade from that mean.

Because the sum of the deviations (signs neglected) is a minimum when taken around the median, the mean deviation is sometimes computed in relation to the median. In practice, however, the mean is generally used and, if the series is symmetrical, the resulting AD is the same.

---

[1] The computation of AD shown in Table 44 may be abbreviated somewhat by the use of a short method, which is described in R. E. Chaddock, *Principles and Methods of Statistics*, pp. 156–158, Houghton Mifflin Co., Boston, 1925 Because of the rather infrequent use of AD the procedure is not given in this text The reader is cautioned that, to use this short method, the assumed mean must be taken as the mid-value of the group in which the mean occurs, or, when the mean falls between two groups (as for the midshipmen's grades), as the mid-value of either of those groups.

The standard deviation, ungrouped data.   Instead of merely neglecting the signs of the deviations from the arithmetic mean, we may square the deviations, thereby making all of them positive.   Thus we may have a measure

$$\sigma^2 = \frac{\Sigma x^2}{N},$$

the *variance* or mean square deviation.   (At a later point we shall use the term *variation* to refer to Σx^2.) σ^2 is also known as the second moment, π_2, of the distribution, since the deviations have been raised to the second power.   We shall make use of the variance in later sections of the book.

At this point we are interested in the square root of this measure,

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}},$$

TABLE 45

COMPUTATION OF STANDARD DEVIATION BY LONG METHOD
FOR SCORES OF 15 PERSONS IN RECALLING TRADE
NAMES OF ADVERTISED PRODUCTS

Subject	Score X	x	x^2
1	12	−20 87	435.56
2	21	−11 87	140 90
3	21	−11 87	140.90
4	23	− 9 87	97.42
5	27	− 5.87	34 46
6	28	− 4.87	23 72
7	30	− 2 87	8 24
8	34	1.13	1.28
9	37	4 13	17.06
10	39	6.13	37.58
11	39	6 13	37.58
12	39	6 13	37.58
13	40	7 13	50.84
14	49	16.13	260.18
15	54	21.13	446 48
Total	493	.	1,769 78

Source: S. M Newhall and M H Heim, "Memory Value of Absolute Size in Magazine Advertising," *Journal of Applied Psychology*, Vol 13, 1929, pp 62–75   The above data were for advertisements of 150 square inches each, and each was observed for 5 seconds   Maximum possible score was 81

$$\overline{X} = \frac{493}{15} = 32 87.$$

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{\frac{1,769\ 78}{15}} = \sqrt{117.98} = 10.9$$

which is termed the *standard deviation* or, occasionally, the root-mean-square deviation. It has been pointed out previously that Σx^2 is a minimum when taken around the arithmetic mean and that the standard deviation is always computed in reference to the arithmetic mean. As the above expression indicates, the steps involved in computing σ are:

(1) Determine the deviation x of each item from $\overline{X}$.
(2) Square these deviations.
(3) Total them.
(4) Divide this sum by N.
(5) Take the square root.

The computation of σ for a series of ungrouped data is shown in Table 45. This procedure involves the computation of x for every item, and would be a rather laborious procedure if there were an appreciably larger number

TABLE 46

COMPUTATION OF STANDARD DEVIATION BY SHORT
METHOD FOR SCORES OF 15 PERSONS IN RECALLING
TRADE NAMES OF ADVERTISED PRODUCTS

Subject	Score X	X^2
1	12	144
2	21	441
3	21	441
4	23	529
5	27	729
6	28	784
7	30	900
8	34	1,156
9	37	1,369
10	39	1,521
11	39	1,521
12	39	1,521
13	40	1,600
14	49	2,401
15	54	2,916
Total	493	17,973

Source Same as Table 45

$$\sigma = \sqrt{\frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2} = \sqrt{\frac{17,973}{15} - \left(\frac{493}{15}\right)^2}$$

$$= \sqrt{1,198.20 - 1,080.22} = \sqrt{117.98}$$

$$= 10.9.$$

of items.   The value of σ may be obtained, without computing each x, by means of the expression[2]

$$\sigma = \sqrt{\frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2}.$$

Referring to Table 45, it will be observed that the value of $\overline{X}$ was rounded to two decimals and thus each value of x and x^2 is an approximation.   If $\overline{X}$ and x are shown to sufficient digits, results by the two methods will be the same.   Here, both methods yield 10.9.

The computation of σ by the short method is illustrated in Table 46.

Notice that the correction $\left(\dfrac{\Sigma X}{N}\right)^2$ is subtracted.   This is always true. The sum of the squared deviations is least when taken around $\overline{X}$.   We, however, took our deviations around some other value (0 in this instance) and these squared deviations are therefore too large.

The standard deviation, grouped data.   Before considering the properties of σ, let us see how to compute σ for a frequency distribution. Since frequencies are present

$$\sigma = \sqrt{\frac{\Sigma f x^2}{N}},$$

where x now represents the deviation of a class mid-value from the mean. Table 47 illustrates the computation of σ for the midshipmen's grades. It is fairly obvious that this method, involving the determination of a number of x values is cumbersome.   It is very unusual for the x values to be integers as in Table 47.   In this case it is because $\overline{X}$ happened to be a class limit.   Consequently our illustration in Table 47 does not show the disadvantages of this method as clearly as might be desired.   If the column of x values had two decimals, the cumbersome nature of this procedure would be clearer.

A short method for σ is available which allows us to take the mid-value of any class as the assumed mean, work with deviations around this value, and make the necessary correction.   The expression is

$$\sigma = \sqrt{\frac{\Sigma f d^2}{N} - \left(\frac{\Sigma f d}{N}\right)^2}.$$

To further shorten the process the deviations are taken in terms of classes. The expression[3] is

$$\sigma = i\sqrt{\frac{\Sigma f(d')^2}{N} - \left(\frac{\Sigma f d'}{N}\right)^2},$$

---

[2] For proof of this expression, see Appendix B, section X–1.
[3] For demonstration, see Appendix B, section X–1.

where d' indicates the deviation of a class mid-value from the assumed mean in terms of classes and i is the class interval. It is of interest to note that the correction factor $\left(\dfrac{\Sigma fd'}{N}\right)^2$ is the square of the correction factor used in computing the arithmetic mean by the short method. The computation of σ by this shorter procedure is shown in Table 48.

TABLE 47

COMPUTATION OF STANDARD DEVIATION BY LONG METHOD FOR GRADES OF THE 1937 GRADUATING CLASS OF THE UNITED STATES NAVAL ACADEMY

$(\overline{X} = 77\ 95)$

Grade	Mid-values of classes X	Number of midshipmen f	Deviation of mid-values from $\overline{X}$ x	fx	fx^2
68 0–69 9	68 95	4	−9 00	− 36	324
70 0–71 9	70 95	17	−7.00	−119	833
72 0–73 9	72 95	39	−5 00	−195	975
74 0–75 9	74 95	62	−3 00	−186	558
76.0–77 9	76 95	58	−1 00	− 58	58
78 0–79 9	78 95	52	1 00	52	52
80 0–81 9	80 95	35	3 00	105	315
82.0–83 9	82.95	22	5 00	110	550
84 0–85 9	84 95	18	7 00	126	882
86 0–87 9	86 95	13	9 00	117	1,053
88 0–89.9	88 95	4	11 00	44	484
90 0–91 9	90 95	2	13 00	26	338
92 0–93 9	92.95	1	15 00	15	225
Total	. .	327	. .	. .	6,647

$$\sigma = \sqrt{\frac{\Sigma fx^2}{N}} = \sqrt{\frac{6,647}{327}} = \sqrt{20.33} = 4.51$$

Properties of the standard deviation. Of the various measures of absolute dispersion which have been mentioned the standard deviation (and its square, the variance) is by far the most important. It will be used in connection with various statistical methods described hereafter. One important consideration is that it is one of the factors involved in the equation for the normal curve and for various skewed curves, discussed in the following chapter. It is also used in testing the reliability of certain statistical measures, in correlation, and in connection with business cycle analysis.

The standard deviation is a much used measure of the spread of a series

of data. If $\pm\sigma$ is measured from the arithmetic mean of a normal distribution, 68.27 per cent of the items are included; within the range of $\overline{X} \pm 2\sigma$, 95.45 per cent are included; and within $\overline{X} \pm 3\sigma$ 99.73 per cent,[4] or nearly all, of the items are included. Chart 106 illustrates what has

TABLE 48

COMPUTATION OF STANDARD DEVIATION BY SHORT METHOD FOR GRADES OF THE 1937 GRADUATING CLASS OF THE UNITED STATES NAVAL ACADEMY

Grade	Number of midshipmen f	d'	fd'	$f(d')^2$
68 0–69 9	4	−5	− 20	100
70 0–71 9	17	−4	− 68	272
72 0–73 9	39	−3	−117	351
74 0–75 9	62	−2	−124	248
76 0–77 9	58	−1	− 58	58
78 0–79 9	52	0	0	0
80 0–81.9	35	1	35	35
82 0–83 9	22	2	44	88
84 0–85.9	18	3	54	162
86 0–87 9	13	4	52	208
88 0–89 9	4	5	20	100
90.0–91 9	2	6	12	72
92.0–93.9	1	7	7	49
Total	327	...	−163	1,743

$$\sigma = i \sqrt{\frac{\Sigma f(d')^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} = 2 0 \sqrt{\frac{1,743}{327} - \left(\frac{-163}{327}\right)^2}$$

$$= 2.0 \sqrt{5\ 0818} = 4\ 51.$$

just been said. The percentages just given refer to a curve of the normal type. If the distribution is skewed, these percentages will be only approximately realized. For the midshipmen's grades (Table 48), $\overline{X} \pm \sigma$ is 77.95 $\pm$ 4.51, or 73.44 to 82.46. Using the procedure described for Q and AD, we find, interpolating into the proper classes, that within these limits are found 222.5 of 327 grades, or 68.04 per cent of the total. Within $\overline{X} \pm 2\sigma$ (that is, from 68.93 to 86.97), we find 311.7, or 95.32 per cent, of the grades. The range $\overline{X} \pm 3\sigma$ runs from 64.42 to 91.48, and includes 325.5, or 99.54 per cent, of the grades. Even though this distribution is skewed, it is apparent that there is substantial agreement with the proportions expected for a normal distribution. It may be observed that the

---

[4] See Appendix E, which gives the areas of one-half of the normal curve (68 27 is twice 34.13447; 95.45 is twice 44.72499; 99.73 is twice 49.86501).

lower value of the $\overline{X} \pm 3\sigma$ range extends below the actual lower limit of the data.   This may occur when a distribution is skewed to the right as is this one.

In dealing with the normal curve in later chapters, we shall not confine ourselves to the proportionate areas included within $\pm\sigma$, $\pm 2\sigma$, and $\pm 3\sigma$ of the mean, but shall consider any desired values of σ.   Thus it may be seen from Appendix E (which gives areas of one-half of a normal curve) that 50 per cent of the items, or area of the curve, would be within $\pm .6745\sigma$
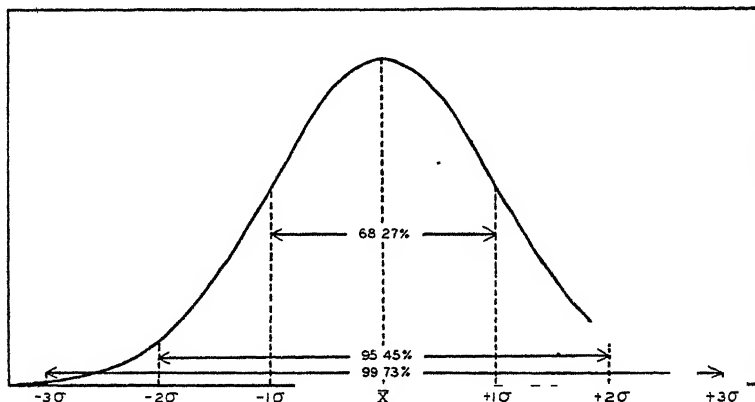
Chart 106.   Proportion of Items Included Within $\pm 1\sigma$, $\pm 2\sigma$, and $\pm 3\sigma$ of the Arithmetic Mean in a Normal Curve.

of the mean and that 90 per cent of the items would be within $\pm 1.645\sigma$ of the mean.

The standard deviation measures the dispersion of a series; the greater the spread of the series, the greater the value of σ.   As a measure of uniformity of the characteristic measured, the smaller the value of σ, the greater the uniformity.   To avoid this inverse relationship, a modification referred to as a measure of precision is sometimes used, especially with reference to the precision of a series of physical measurements.   This measure is

$$h^2 = \frac{1}{2\sigma^2}.$$

It is not often used in statistical work in the social sciences.

The value of σ which we have computed is a measure of the dispersion of the data (usually a sample) upon which it is based.   It is not a measure, or an estimate, of the dispersion of the population from which the sample was drawn.   The problem of estimating the dispersion of the population will be considered in Chapter XII.

Comparison of measures of absolute dispersion. We have considered the range, the quartile deviation, the mean deviation, and the standard deviation. It was seen that the range is useful only for the crudest sort of statement. The quartile deviation and mean deviation, while occasionally serviceable, are not so valuable to us as the standard deviation, which appears as a necessary or useful element in curve fitting, testing the reliability of means and other measures, time series analysis, and correlation.

In the case of a normal distribution it was pointed out that 50 per cent of the items were present within $\pm 1Q$ of the median (or mean), 57.5 per cent within $\pm 1AD$ of the mean, and 68.27 per cent within $\pm 1\sigma$ of the mean. For a normal distribution these measures consequently bear fixed relations to each other, as follows:[5]

$$Q = .6745\ \sigma.$$
$$AD = .7979\ \sigma.$$
$$Q = .8453\ AD.$$
$$\sigma = 1.2533\ AD.$$
$$AD = 1.1830\ Q.$$
$$\sigma = 1.4826\ Q.$$

Measures of Relative Dispersion

In the preceding paragraphs we have discussed measures of absolute dispersion, all of which are expressed in terms of the units of the problem, which may be dollars, pounds, inches, percentages, etc. When we wish to compare the dispersions of two or more series, it may or may not be desirable to use such a measure. The comparison of dispersions of two or more series resolves itself into three possible situations.

(1) The series to be compared may be expressed in the same units, and the means may be the same, or nearly the same, in size. The grades of the midshipmen showed a mean of 77.95 and a standard deviation of 4.51. If another Naval Academy class showed $\overline{X} = 78.01$ and $\sigma = 3.75$, it is clear that the second class would exhibit less dispersion.

(2) The series to be compared may be expressed in the same units, but the arithmetic means may differ. Some years ago the Goodyear Tire and Rubber Company developed a new type of cord for automobile tires which was designated as "Supertwist." The Supertwist cord was superior to ordinary cord in that it could stretch more and had a longer flex life.

---

[5] The first two relationships may be obtained by interpolating into Appendix E. The others are computed from them.

Tests made on cord as received from the cotton mill and prior to fabrication into tires showed for the flex life of Supertwist cord

$$\overline{X} = 138.64 \text{ minutes, and } \sigma = 15.27 \text{ minutes;}$$

while for regular cord the figures were

$$\overline{X} = 87.66 \text{ minutes, and } \sigma = 14.12 \text{ minutes.}$$

If we compare the two σ values, it appears that Supertwist cord is more variable in respect to flex life than is regular cord. However, it must be noted that the average flex life of Supertwist is much greater than that of regular cord. Taking this factor into consideration, we may set up a measure of *relative dispersion,*

$$V = \frac{\sigma}{\overline{X}}.$$

This is the coefficient of variation and is usually expressed as a percentage. For the Supertwist cord

$$V = \frac{15.27}{138.64} = 0\ 1101, \text{ or } 11.0 \text{ per cent;}$$

while for regular cord

$$V = \frac{14.12}{87.66} = 0.1611, \text{ or } 16.1 \text{ per cent.}$$

It is thus apparent that the relative variation in flex life is much less for Supertwist cord than for regular cord.

Chart 107 also illustrates the comparison of dispersions of two series having different mean values. In section A are shown the curves of two distributions having the same absolute dispersions but different relative dispersions. In section B are curves of two distributions having quite different absolute dispersions but the same relative dispersions. If the zero is shown on the horizontal scale, as in Chart 107, a very rough visual impression may be had of the relative dispersion of a series. For this reason some statisticians think it is desirable to show the zero on the horizontal scale. This does not seem to be a very important matter, however, since relative dispersion can at best be visualized only very roughly by this device. Occasionally frequency distributions are formed with class intervals expressed, not in terms of original units, but as percentages of the mean, the interval being some convenient figure, such as 10 per cent of the mean. If two such distributions are plotted on one chart, it is easy to compare visually their relative dispersions.

(3) The series to be compared may be expressed in different units. In such a case the standard deviations cannot be directly compared. A

study of a large number of male industrial workers[6] revealed an average pulse rate of 81.1 beats per minute and a standard deviation of about 12.2 beats per minute.   Measurements of height showed $\overline{X}$ = 66.9 inches and σ = 2.7 inches.   The measurements of height included a small number of men not measured as to pulse rate.   Let us disregard this difficulty for the purposes of our illustration.   Are the industrial workers more variable in respect to pulse rate or height?   It is obvious that the two standard
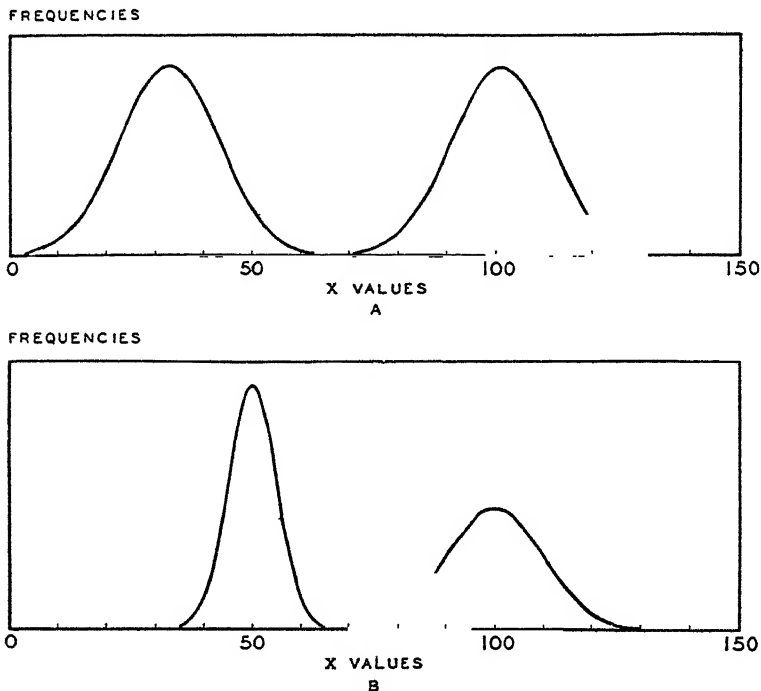
Chart 107.   Comparisons of Dispersions of Series Having Different Arithmetic Means.   A.   Same absolute dispersion, different relative dispersion: left-hand curve, $\overline{X}$ = 33, σ = 10, V = 30.3 per cent; right-hand curve, $\overline{X}$ = 101, σ = 10, V = 9 9 per cent.   B.   Different absolute dispersion, same relative dispersion: left-hand curve, $\overline{X}$ = 50, σ = 5, V = 10 per cent; right-hand curve, $\overline{X}$ = 100, σ = 10, V = 10 per cent. (Sections A and B have different vertical scales since they are not intended to be compared.   However, if the vertical scale of section B is expanded 50 per cent, all curves will have the same area.)

deviations, being in different units, cannot be compared.   Computing the two coefficients of variation shows for pulse rate

$$V = \frac{12.2}{81.1} = .149, \text{ or } 14.9 \text{ per cent,}$$

[6] Based on data in *A Health Study of Ten Thousand Male Industrial Workers*, pp. 45 and 59, United States Public Health Service, Public Health Bulletin, 162.

and for height

$$V = \frac{2.7}{66.9} = .040, \text{ or } 4.0 \text{ per cent.}$$

It is clear that, for this group of men, pulse rate is subject to greater dispersion than is height.

Somewhat akin to our measurement of relative dispersion is the possibility of expressing a given value in terms of its divergence from the mean and also in terms of the dispersion of the series. Such a procedure is not especially useful when we are considering only one value or comparing two values from the same series. Its usefulness becomes apparent when we want to compare two values from different series and when those two series (1) differ in respect to $\overline{X}$ or σ, or both, or (2) are expressed in different units. Suppose that a certain student has made a grade of 180 on an intelligence test, and that his group showed $\overline{X} = 160$ and $\sigma = 15$. This same student made a grade of 86 in history, and the group showed $\overline{X} = 70$ and $\sigma = 12$. We are interested in knowing whether his relative standing is higher in the intelligence test or in history. In the intelligence test he was 20 points above the mean, and in history he was 16 points above the mean. These deviations, however, are not comparable, but may be rendered so by dividing by their respective standard deviations. Thus

Intelligence test: $\dfrac{X - \overline{X}}{\sigma} = \dfrac{180 - 160}{15} = \dfrac{+20}{15} = +1.33;$

History: $\dfrac{X - \overline{X}}{\sigma} = \dfrac{86 - 70}{12} = \dfrac{+16}{12} = +1.33.$

It is apparent that the student shows the same relative standing in each group, being $+1.33\ \sigma$ above the mean in each. The usefulness of this device is by no means limited to the educational field. It is, however, often used with test data and is then referred to as a "standard score."

Skewness

When a series is not symmetrical, it is said to be asymmetrical or skewed. In Chart 103 we showed such a skewed curve in relation to a symmetrical one. The curve of midshipmen's grades (Chart 108) is also skewed. Our measures of skewness indicate not only the amount of skewness but also the direction. A series is said to be skewed in the direction of the extreme values, or, speaking in terms of the curve, in the direction of the excess tail. Thus the two curves referred to above are both skewed positively, or to the right. Most skewed curves encountered in the social sciences are skewed to the right. Only rarely do we find curves skewed

NUMBER OF
MIDSHIPMEN

MED = 77.38

MO = 75 65 $\bar{X}$ = 77 95
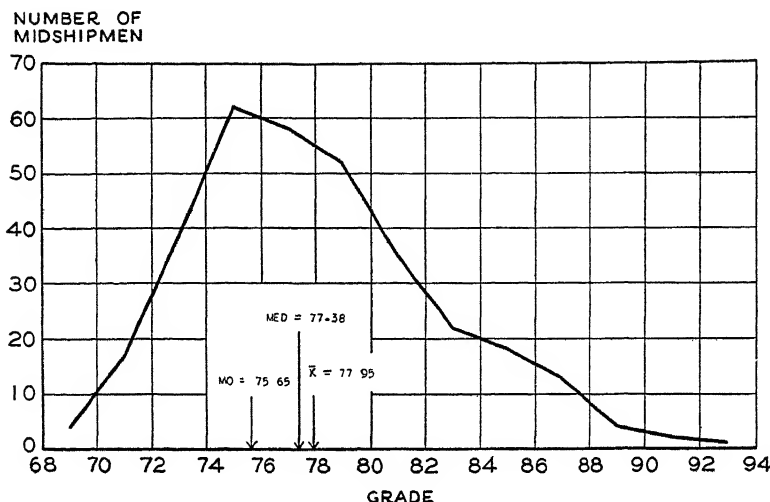
GRADE

Chart 108.   Grades of the 1937 Graduating Class of the United States Naval Academy,
Showing Location of Mean, Median, and Mode.

to the left, such as those shown in Charts 85 and 109, and even less rarely
do we find data *characteristically* skewed to the left.

Many series, however, are characteristically skewed to the right.   Ex-
amples are frequency distributions of wages or salaries, use of electricity
(see Chart 125, p. 293), weights of adult male human beings, and numer-
ous other variables.   Distributions of grades are apt to be moderately

NUMBER OF
INVENTORS

AGE IN YEARS
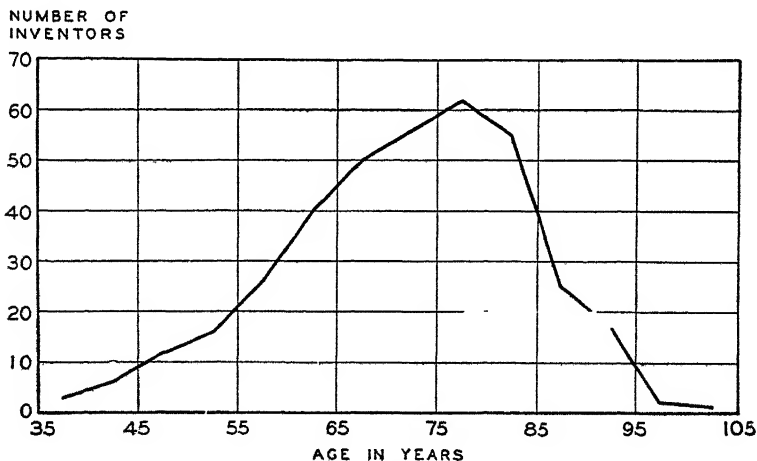
Chart 109.   Age at Death of 371 American Inventors.   (Data from "Bio-Social
Characteristics of American Inventors," by Sanford Winston, *American Sociological
Review*, Vol. 2, No. 6, December 1937, pp 837–849.)

skewed to the right, or nearly symmetrical. In the case of the midshipmen's grades the skewness is partly due to the fact that we are considering only those men who had survived the previous three years, during which some of the less able had been dropped. The distribution of ages at death of the American inventors in Chart 109 may be characteristically skewed to the left, or the skewness may be due to the fact that a time factor is present—almost one-fifth of the inventors included in this study were born before 1800.

Pearsonian measure of skewness. It was pointed out in the preceding chapter that the mode is not influenced by the presence of extreme values, the median is influenced by their position only, and the arithmetic mean is influenced by the size of the extremes. Consequently we could make use of the mode and the mean to measure skewness. We might say then that skewness = mean − mode. But there are some shortcomings of such a measure. In the first place it is a measure of absolute skewness and would have much different meaning for a series of small dispersion than for a widely dispersed series. It is rather unusual for two (or more) series to have the same σ and therefore we practically never use a measure of absolute skewness, preferring a measure of relative skewness. The measure just mentioned may be put into relative terms by dividing by σ. Now

$$\text{Skewness} = \frac{\overline{X} - \text{Mo}}{\sigma}.$$

This gives us a relative measure with positive sign when skewness is to the right, and with negative sign when skewness is to the left. There is, however, another important difficulty growing out of the fact that the mode is not very satisfactorily located for most frequency distributions. The median is rather satisfactorily located and therefore we use the measure[7]

$$\text{Sk} = \frac{3(\overline{X} - \text{Med})}{\sigma}.$$

In the preceding chapter we found that, for the midshipmen's grades, the value of $\overline{X}$ was 77.95, while the median was 77.38. In this chapter the value of σ was found to be 4.51. The skewness, then, is

$$\text{Sk} = \frac{3(77.95 - 77.38)}{4.51} = +.38.$$

[7] The presence of the 3 in the expression is explained as follows: Karl Pearson showed empirically that in moderately skewed distributions of a continuous variable the median fell about ⅔ of the distance from the mode toward the mean. Consequently we have $\text{Mo} = \overline{X} - 3(\overline{X} - \text{Med})$ Now if we substitute this expression for the mode in the measure of skewness, we have

$$\text{Sk} = \frac{\overline{X} - [\overline{X} - 3(\overline{X} - \text{Med})]}{\sigma} = \frac{3(\overline{X} - \text{Med})}{\sigma}.$$

TABLE 49

Computation of Various Measures for Age at Death of 371 American Inventors

Age at death in years	f	d'	fd'	$f(d')^2$	$f(d')^3$
35 and under 40	3	-6	-18	108	-648
40 and under 45	6	-5	-30	150	-750
45 and under 50	12	-4	-48	192	-768
50 and under 55	16	-3	-48	144	-432
55 and under 60	26	-2	-52	104	-208
60 and under 65	40	-1	-40	40	-40
65 and under 70	50	0	0	0	0
70 and under 75	56	1	56	56	56
75 and under 80	62	2	124	248	496
80 and under 85	55	3	165	495	1,485
85 and under 90	25	4	100	400	1,600
90 and under 95	17	5	85	425	2,125
95 and under 100	2	6	12	72	432
100 and over*	1	7	7	49	343
Total... ...	371	...	$+313$	2,483	$+3,691$

* This class assumed to have its mid-value at 102 5

Source Sanford Winston, "Bio-social Characteristics of American Inventors," *American Sociological Review*, Vol 2, No 6, December 1937, p 848 and by correspondence

$$\frac{N}{2} = 185.5. \qquad \frac{N}{4} = 92.75. \qquad \frac{N}{10} = 37.1.$$

$$Q_1 = 60 + \frac{29\ 75}{40} \times 5 = 63.72 \text{ years.}$$

$$Q_3 = 85 - \frac{47.75}{55} \times 5 = 80.66 \text{ years.}$$

$$\text{Median} = 70 + \frac{32.5}{56} \times 5 = 72.90 \text{ years.} \qquad \overline{X} = 67.5 + \frac{313}{371} \times 5 = 71.72 \text{ years}$$

$$P_{10} = 55 + \frac{0\ 1}{26} \times 5 = 55.02 \text{ years.}$$

$$P_{90} = 90 - \frac{17.1}{25} \times 5 = 86.58 \text{ years.}$$

$$\sigma = 5 \sqrt{\frac{2,483}{371} - \left(\frac{313}{371}\right)^2} = 12.23 \text{ years.}$$

$$\nu_1 = \frac{\Sigma fd'}{N} = \frac{+313}{371} = .843666.$$

$$\nu_2 = \frac{\Sigma f(d')^2}{N} = \frac{2,483}{371} = 6.692722.$$

$$\nu_3 = \frac{\Sigma f(d')^3}{N} = \frac{+3,691}{371} = 9.948787.$$

$$\pi_1 = 0.$$

$$\pi_2 = \nu_2 - \nu_1^2 = 6.692722 - (843666)^2 = 5.980950.$$

$$\pi_3 = \nu_3 - 3\nu_1\nu_2 + 2\nu_1^3 = +9.948787 - 3(.843666)(6.692722) + 2(.843666)^3$$
$$= -5.789483.$$

This may be considered as a moderate degree of skewness since the measure varies within the limits[5] of ±3. It should be added that values as large as ±1 are rather unusual.

For the data of age at death of the American inventors, it is shown under Table 49 that $\overline{X} = 71.72$ years, while Med = 72.90 years and $\sigma = 12.23$ years. The Pearsonian measure of skewness is

$$\text{Sk} = \frac{3(71.72 - 72.90)}{12.23} = -.29.$$

Measures of skewness based on quartiles and percentiles. It was previously pointed out that in a symmetrical distribution Q_1 and Q_3 lie equidistant from the median. If a series is skewed to the right, Q_3 is farther from the median than is Q_1. If a series is skewed to the left, Q_1 is farther from the median (Q_2) than is Q_3. An absolute measure of skewness, then, may be given by

$$(Q_3 - Q_2) - (Q_2 - Q_1) = Q_1 + Q_3 - 2Q_2.$$

This measure is put into relative terms by dividing by the quartile deviation. Thus as a measure of skewness we could use

$$\frac{Q_1 + Q_3 - 2Q_2}{\frac{Q_3 - Q_1}{2}}.$$

It can be shown[9] that this value varies within the limits of ±2. It is perhaps a little more satisfactory to use as a measure of skewness

$$\text{Sk}_Q = \frac{Q_1 + Q_3 - 2Q_2}{Q_3 - Q_1},$$

which, of course, varies within the limits of ±1. We may call this the "quartile measure of skewness."

For the grades of the midshipmen, the values of Q_1, Q_3, and the median were previously obtained. Thus

$$\text{Sk}_Q = \frac{74.65 + 80.71 - 2(77.38)}{80.71 - 74.65} = +.10.$$

For the age at death of the American inventors, the values of the quartiles and the medians are shown below Table 49. The skewness, then, is

$$\text{Sk}_Q = \frac{80.66 + 63.72 - 2(72.90)}{80.66 - 63.72} = -.08.$$

---

[8] Harold Hotelling and Leonard M. Solomons ("The Limits of a Measure of Skewness," *Annals of Mathematical Statistics*, May 1932, pp. 141–142) have shown that $\dfrac{\overline{X} - \text{Med}}{\sigma}$ lies between ±1.

[9] See Appendix B, section X-2. The maximum value in actual use will, of course, be less than 2.

It is apparent that the upper 25 per cent and the lower 25 per cent of the data have only a positional influence in determining the value of this measure. All the items in the upper quarter might be closely clustered or they might be spread out over a wide range, yet their influence on Sk_Q would be the same.

A somewhat more sensitive measure of skewness may be based upon the deciles or percentiles. We could thus consider the 10th and 90th percentiles in relation to the 50th percentile (the median). The measure of absolute skewness would be $(P_{90} - P_{50}) - (P_{50} - P_{10})$, or $P_{90} + P_{10} - 2P_{50}$. If this measure is divided by the 10–90 percentile range (that is, $P_{90} - P_{10}$), we have a measure of relative skewness

$$Sk_P = \frac{P_{10} + P_{90} - 2P_{50}}{P_{90} - P_{10}},$$

which varies within the limits[10] of ± 1.

Computing Sk_P for the midshipmen's grades, we refer to the earlier part of this chapter for P_{10} and P_{90}, which have previously been computed. These were $P_{10} = 72.55$ and $P_{90} = 84.54$. For the measure of skewness, then

$$Sk_P = \frac{72.55 + 84.54 - 2(77.38)}{84.54 - 72.55} = +.19.$$

The values of P_{10} and P_{90} for the age at death of the inventors are shown below Table 49. Computing the skewness gives

$$Sk_P = \frac{86.58 + 55.02 - 2(72.90)}{86.58 - 55.02} = -.13.$$

Measure of skewness based on the third moment. We have seen that the most satisfactory measure of dispersion is the standard deviation, which is based upon the second moment about the mean

$$\pi_2 = \frac{\Sigma x^2}{N}, \text{ and } \sigma = \sqrt{\pi_2} = \sqrt{\frac{\Sigma x^2}{N}}.$$

A very useful measure of skewness may be obtained by making use of the third moment about the mean

$$\pi_3 = \frac{\Sigma x^3}{N}.$$

It will be recalled that the first moment about the mean

$$\pi_1 = \frac{\Sigma x}{N}$$

---

[10] The proof of this exactly parallels that given in Appendix B, section X-2. The maximum value in actual use will, of course, be appreciably less than 1.

is always zero. However, the third moment about the mean is not zero unless the distribution is symmetrical about the mean. Cubing a deviation does not change its sign. It does, however, have a disproportionately large effect on large deviations. As illustrations, consider the two sets of data given in Tables 50 and 51, the first of which is symmetrical around a mean of 6, while the second is not symmetrical around a mean of 6. Both sets of data have

$$\pi_1 = \frac{\Sigma x}{N} = 0,$$

and the data of Table 50 have

$$\pi_3 = \frac{\Sigma x^3}{N} = 0.$$

But the figures in Table 51 show

$$\pi_3 = \frac{\Sigma x^3}{N} = +6.$$

TABLE 50	TABLE 51
COMPUTATION OF FIRST AND THIRD MOMENTS OF A SYMMETRICAL SERIES	COMPUTATION OF FIRST AND THIRD MOMENTS OF AN ASYMMETRICAL SERIES

X	x	x^3	X	x	x^3
2	-4	-64	3	-3	-27
4	-2	$- 8$	4	-2	$- 8$
6	0	0	6	0	0
8	$+2$	$+ 8$	7	$+1$	$+ 1$
10	$+4$	$+64$	10	$+4$	$+64$
	0	0		0	$+30$

$$\pi_1 = \frac{\Sigma x}{N} = \frac{0}{5} = 0. \qquad \pi_1 = \frac{\Sigma x}{N} = \frac{0}{5} = 0.$$

$$\pi_3 = \frac{\Sigma x^3}{N} = \frac{0}{5} = 0. \qquad \pi_3 = \frac{\Sigma x^3}{N} = \frac{+30}{5} = +6.$$

To compute the third moment of a frequency distribution,

$$\pi_3 = \frac{\Sigma f x^3}{N},$$

taking the actual deviations from the arithmetic mean, cubing them, multiplying by the frequencies, summing, and dividing by N, would be laborious. As shown in Appendix B (section X–1), the second moment σ^2 or π_2 can be obtained by a short process. In terms of class intervals,

$$\pi_2 = \frac{\Sigma f(d')^2}{N} - \left(\frac{\Sigma f d'}{N}\right)^2$$

The value of the third moment (also in terms of class intervals) is given by[11]

$$\pi_3 = \frac{\Sigma f(d')^3}{N} - 3\frac{\Sigma fd'}{N}\frac{\Sigma f(d')^2}{N} + 2\left(\frac{\Sigma fd'}{N}\right)^3$$

Or, letting $\nu_1 = \dfrac{\Sigma fd'}{N}$, $\nu_2 = \dfrac{\Sigma f(d')^2}{N}$, and $\nu_3 = \dfrac{\Sigma f(d')^3}{N}$,

$$\pi_2 = \nu_2 - \nu_1^2,$$

and

$$\pi_3 = \nu_3 - 3\nu_1\nu_2 + 2\nu_1^3.$$

Obviously, π_3 is a measure of absolute skewness. It is put into relative terms by dividing by σ^3. The measure of relative skewness, based on the third moment,[12] is

$$\alpha_3 = \frac{\pi_3}{\sigma^3} \text{ or } \frac{\pi_3}{\sqrt{\pi_2^3}},$$

where σ is in class intervals.
The symbol $\sqrt{\beta_1}$ is also used to identify this measure.

The values of the first, second, and third moments for the data of mid shipmen's grades are computed in Table 52. From these we obtain

$$\alpha_3 = \frac{\pi_3}{\sqrt{\pi_2^3}} = \frac{+6.527531}{\sqrt{(5.081802)^3}} = +.57.$$

Similarly, the first three moments for the age at death of the American inventors have been computed in Table 49, showing

$$\alpha_3 = \frac{-5.789483}{\sqrt{(5.980950)^3}} = -.40.$$

If the distribution is symmetrical, α_3, of course, is zero. Sometimes α_3^2 or β_1 is used as a measure of skewness and

$$\alpha_3^2 = \beta_1 = \frac{\pi_3^2}{\pi_2^3}.$$

No definite upper limit is apparent for α_3 or β_1, but values as great as ± 2 for α_3 indicate marked skewness.

---

[11] See Appendix B, section X-3.
[12] No previous mention has been made of α_1 or α_2. For any series of figures,

$$\alpha_1 = \frac{\pi_1}{\sigma} \text{ or } \frac{\pi_1}{\sqrt{\pi_2}} = 0;$$

$$\alpha_2 = \frac{\pi_2}{\sigma^2} \text{ or } \frac{\pi_2}{\sqrt{\pi_2^2}} = 1.$$

TABLE 52

COMPUTATION OF FIRST THREE MOMENTS FOR GRADES OF THE 1937 GRADUATING
CLASS OF THE UNITED STATES NAVAL ACADEMY

Grade	Number of midshipmen f	d'	fd'	$f(d')^2$	$f(d')^3$
68 0–69 9	4	−5	− 20	100	− 500
70.0–71.9	17	−4	− 68	272	−1,088
72.0–73.9	39	−3	−117	351	−1,053
74.0–75.9	62	−2	−124	248	− 496
76.0–77.9	58	−1	− 58	58	− 58
78 0–79.9	52	0	0	0	0
80 0–81.9	35	1	35	35	35
82.0–83.9	22	2	44	88	176
84.0–85.9	18	3	54	162	486
86.0–87 9	13	4	52	208	832
88.0–89.9	4	5	20	100	500
90.0–91.9	2	6	12	72	432
92.0–93.9	1	7	7	49	343
Total	327	..	−163	1,743	−391

$$\nu_1 = \frac{\Sigma fd'}{N} = \frac{-163}{327} = -.498471$$

$$\nu_2 = \frac{\Sigma f(d')^2}{N} = \frac{1,743}{327} = 5.330275.$$

$$\nu_3 = \frac{\Sigma f(d')^3}{N} = \frac{-391}{327} = -1.195719.$$

$\pi_1 = 0.$

$\pi_2 = \nu_2 - \nu_1^2 = 5.330275 - (-.498471)^2 = 5.081802.$

$\pi_3 = \nu_3 - 3\nu_1\nu_2 + 2\nu_1^3 = -1.195719 - 3(-.498471)(5.330275) + 2(-.498471)^3$
$= +6.527531.$

In the preceding chapter it was pointed out in footnote 5 that the value of the mode may be estimated by use of the expression

$$Mo = \overline{X} - \sigma\frac{\sqrt{\beta_1}(\beta_2 + 3)}{2(5\beta_2 - 6\beta_1 - 9)}.$$

The final fraction of this expression is sometimes used as a measure of skewness. It is often called χ (chi) but, since we use χ^2 in a different sense in the following chapter, we shall refer to it as Sk_β and

$$Sk_\beta = \frac{\sqrt{\beta_1}(\beta_2 + 3)}{2(5\beta_2 - 6\beta_1 - 9)}.$$

The computation of β_2 is described in the following section. The sign for this measure of skewness is obtained by giving Sk_β the same sign as π_3.

Kurtosis

Chart 110 shows a distribution which is more peaked than normal. Such a distribution is referred to as *leptokurtic*. A *platykurtic*, or flat-topped, distribution is shown in Chart 111. The normal curve is desig-

NUMBER
OF HOUSES
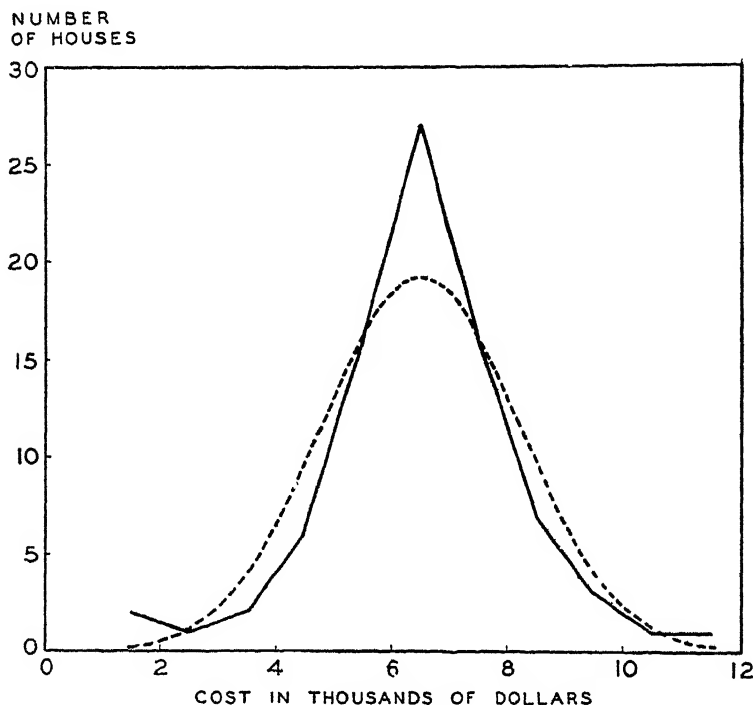
Chart 110. Cost of New 5-Room House and Lot to Purchaser, Cleveland, 1924, and Normal Curve Having Same N, $\overline{X}$, and σ. (Based on data of Table 53.)

nated as *mesokurtic*.[13] The degree of kurtosis present in a series may be measured by making use of the fourth moment,

$$\pi_4 = \frac{\Sigma x^4}{N},$$

or, for a frequency distribution,

$$\pi_4 = \frac{\Sigma f x^4}{N}.$$

---

[13] *Kurtic* = humpbacked; thus humped or unimodal. *Lepto* = slender, narrow. *Platy* = broad, wide, flat. *Meso* = in the middle, intermediate

By a procedure similar to that given in Appendix B, section X–3, it may be shown that

$$\pi_4 = \frac{\Sigma f(d')^4}{N} - 4\frac{\Sigma fd'}{N}\frac{\Sigma f(d')^3}{N} + 6\left(\frac{\Sigma fd'}{N}\right)^2\frac{\Sigma f(d')^2}{N} - 3\left(\frac{\Sigma fd'}{N}\right)^4.$$

or letting

$$\nu_4 = \frac{\Sigma f(d')^4}{N}$$

$$\pi_4 = \nu_4 - 4\nu_1\nu_3 + 6\nu_1^2\nu_2 - 3\nu_1^4.$$

Now π_4 gives an absolute expression for kurtosis. This may be put into relative terms by dividing by $\sigma^4 = \pi_2^2$. The measure is known as α_4 or β_2, and

$$\alpha_4 = \beta_2 = \frac{\pi_4}{\sigma^4}, \text{ or } \frac{\pi_4}{\sqrt{\pi_2^4}} = \frac{\pi_4}{\pi_2^2},$$

where σ is in class intervals.
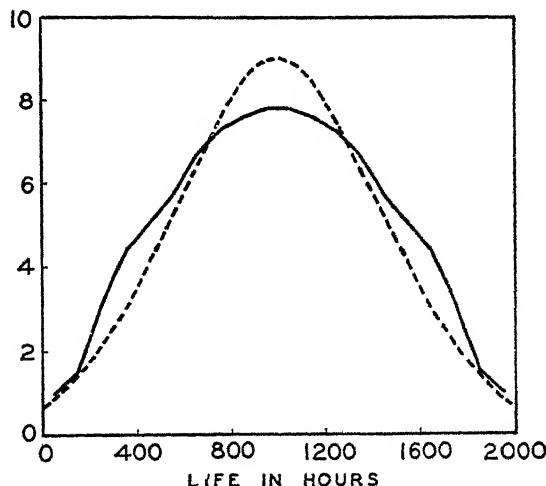
Chart 111. Length of Life of a Group of Electric Lamps and Normal Curve Having Same N, $\overline{X}$, and σ. (Based on data of Table 54. The tails of the normal curve are not shown. The left tail would cross the Y axis.)

This expression has a value of 3.0 for the normal curve. For a flat-topped curve, $\alpha_4 < 3.0$. For a peaked curve, $\alpha_4 > 3.0$.

The peaked curve of Chart 110 is shown in comparison with a normal curve having the same N, $\overline{X}$, and σ. In Table 53 the moments of this leptokurtic distribution have been computed and the value of α_4 or $\beta_2 = 4.46$.

The flat-topped, or platykurtic, curve in Chart 111 is also shown in relation to a normal curve having the same N, $\overline{X}$, and σ. The moments of the flat-topped series are shown in Table 54 and from these α_4 or β_2 is found to be 2.22.

TABLE 53

Computation of First Four Moments for Cost of New 5-Room Wood House and Lot to Purchaser, Cleveland, 1924

Cost (mid-values)	f	d'	fd'	$f(d')^2$	$f(d')^3$	$f(d')^4$
\$ 1,500	2	-5	-10	50	-250	1,250
2,500	1	-4	-4	16	-64	256
3,500	2	-3	-6	18	-54	162
4,500	6	-2	-12	24	-48	96
5,500	16	-1	-16	16	-16	16
6,500	27	0	0	0	0	0
7,500	16	1	16	16	16	16
8,500	7	2	14	28	56	112
9,500	3	3	9	27	81	243
10,500	1	4	4	16	64	256
11,500	1	5	5	25	125	625
Total	82		0	236	-90	3,032

Source Frank R Garfield and William M Hood, "Construction Costs and Real Property Values," *Journal of the American Statistical Association*, Vol 32, No 200, December 1937, p 647. Data are those shown in Chart I for 5-room wood houses

$$\nu_1 = \frac{\Sigma fd'}{N} = \frac{0}{82} = 0.$$

$$\nu_2 = \frac{\Sigma f(d')^2}{N} = \frac{236}{82} = 2\,878049.$$

$$\nu_3 = \frac{\Sigma f(d')^3}{N} = \frac{-90}{82} = -1\,097561.$$

$$\nu_4 = \frac{\Sigma f(d')^4}{N} = \frac{3,032}{82} = 36\,975601.$$

$$\pi_1 = 0.$$
$$\pi_2 = \nu_2 - \nu_1^2 = 2\,878049.$$
$$\pi_3 = \nu_3 - 3\nu_1\nu_2 + 2\nu_1^3 = -1\,097561$$
$$\pi_4 = \nu_4 - 4\nu_1\nu_3 + 6\nu_1^2\nu_2 - 3\nu_1^4 = 36\,975601.$$
$$\alpha_4 = \frac{\pi_4}{\pi_2^2} = \frac{36\,975601}{(2\,878049)^2} = 4\,46.$$

NOTE: The assumed mean (\$6,500) and the mean coincide, resulting in a value of 0 for ν_1. There are therefore no differences between the ν and π values, since $\nu_1^2 = 0$, $\nu_1\nu_2 = 0$, $\nu_1^3 = 0$, $\nu_1\nu_3 = 0$, etc.

When a deviation is raised to a fourth or a second power, its sign becomes positive. The fourth power increases extreme deviations disproportionately in comparison with raising them to the second power. Con-

TABLE 54

COMPUTATION OF FIRST FOUR MOMENTS FOR LENGTH OF LIFE OF A GROUP OF
ELECTRIC LAMPS

Length of life in hours (mid-values)	Percentage frequencies f	d'	fd'	$f(d')^2$	$f(d')^3$	$f(d')^4$
50	1 0	−9	− 9.0	81.0	− 729 0	6,561 0
150	1.5	−8	−12.0	96.0	− 768.0	6,144.0
250	3.1	−7	−21 7	151.9	−1,063.3	7,443.1
350	4.4	−6	−26 4	158.4	− 950 4	5,702.4
450	5 0	−5	−25 0	125.0	− 625.0	3,125 0
550	5 7	−4	−22.8	91.2	− 364.8	1,459 2
650	6.6	−3	−19 8	59 4	− 178.2	534 6
750	7 3	−2	−14.6	29 2	− 58 4	116 8
850	7.6	−1	− 7.6	7 6	− 7 6	7 6
950	7.8	0	0	0	0	0
1050	7.8	1	7 8	7 8	7 8	7 8
1150	7.6	2	15 2	30.4	60.8	121 6
1250	7.3	3	21.9	65 7	197 1	591 3
1350	6.6	4	26 4	105.6	422 4	1,689.6
1450	5 7	5	28 5	142 5	712 5	3,562 5
1550	5.0	6	30 0	180 0	1,080 0	6,480 0
1650	4.4	7	30 8	215 6	1,509.2	10,564 4
1750	3.1	8	24.8	198.4	1,587.2	12,697.6
1850	1.5	9	13 5	121 5	1,093.5	9,841.5
1950	1.0	10	10.0	100 0	1,000.0	10,000.0
Total	100.0	...	+50 0	1,967 2	+2,925.8	86,650.0

Source. Robley Winfrey and Edwin B Kurtz, *Life Characteristics of Physical Property*, Bulletin 103, Iowa Engineering Experiment Station, p. 58, Property Group 28–2.

$$\nu_1 = \frac{\Sigma fd'}{N} = \frac{+50}{100.0} = +.50.$$

$$\nu_2 = \frac{\Sigma f(d')^2}{N} = \frac{1,967\ 2}{100.0} = 19.672.$$

$$\nu_3 = \frac{\Sigma f(d')^3}{N} = \frac{+2,925.8}{100\ 0} = +29.258.$$

$$\nu_4 = \frac{\Sigma f(d')^4}{N} = \frac{86,650\ 0}{100.0} = 866.500.$$

$\pi_1 = 0.$

$\pi_2 = \nu_2 - \nu_1^2 = 19.672 - (.50)^2 = 19.422.$

$\pi_3 = \nu_3 - 3\nu_1\nu_2 + 2\nu_1^3 = 29.258 - 3(.50)(19.672) + 2(.50)^3 = 0.$

$\pi_4 = \nu_4 - 4\nu_1\nu_3 + 6\nu_1^2\nu_2 - 3\nu_1^4 = 866.500 - 4(.50)(29.258) + 6(.50)^2(19.672) - 3(.50)^4$

$\qquad = 837.3045.$

$$\alpha_4 = \frac{\pi_4}{\pi_2^2} = \frac{837\ 3045}{(19.422)^2} = 2.22.$$

sequently the narrower the shoulders of a distribution and the longer the tails, the greater will be π_4 in relation to π_2^2

Correction of the Moments for Grouping Error

Recapitulating what has been said earlier concerning the first four mo ments of a frequency distribution:

Moments around an arbitrary origin:

$$\nu_1 = \frac{\Sigma f d'}{N}.$$

$$\nu_2 = \frac{\Sigma f (d')^2}{N}.$$

$$\nu_3 = \frac{\Sigma f (d')^3}{N}.$$

$$\nu_4 = \frac{\Sigma f (d')^4}{N}.$$

Moments around the arithmetic mean:

$$\pi_1 = 0.$$

$$\pi_2 = \nu_2 - \nu_1^2.$$

$$\pi_3 = \nu_3 - 3\nu_1\nu_2 + 2\nu_1^3.$$

$$\pi_4 = \nu_4 - 4\nu_1\nu_3 + 6\nu_1^2\nu_2 - 3\nu_1^4.$$

In computing the mean, the standard deviation, π_3, and π_4 for frequency distributions, we made use of the mid-values of the classes as representative values. We saw, in the previous chapter, that the mid-values were incorrect assumptions but that the errors present tend to offset each other when we compute the arithmetic mean. This offsetting is also present when the third moment is computed. It will be remembered that the mid-values of the classes preceding the modal class tend to be too small, while the mid-values of the classes following the modal class tend to be too large (see Table 38). The result is that the various x values are slightly larger (in absolute value) than they should be and no offsetting occurs when they are squared or raised to the fourth power. Consequently the value of π_2 (and σ) and the value of π_4 are apt to be slightly larger than the values computed from the same data ungrouped. Sheppard's

corrections attempt to offset this upward bias. The corrected moments are indicated by μ and are:[14]

$$\mu_1 = \pi_1 = 0.$$
$$\mu_2 = \pi_2 - \tfrac{1}{12}.$$
$$\mu_3 = \pi_3.$$
$$\mu_4 = \pi_4 - \tfrac{1}{2}\pi_2 + \tfrac{7}{240}.$$

where all computations are in terms of class intervals.

If we use the class means instead of the mid-values, the arithmetic mean can be computed accurately. However, if class means are used, the values of π_2 (σ^2) and π_4 will be smaller than if computed from the same data ungrouped. We shall give an arithmetic illustration to show that, when the mean of each of several groups of figures is substituted for those figures, σ for the series is decreased; that is, it has a downward bias.

Consider the two following sets of data. The first contains nine different values; the second shows the mean of the first three items repeated three times, the mean of the second three items repeated three times, and the mean of the last three items repeated three times. The standard deviation of the nine different items is 2.58, but the standard deviation of the three groups of means is 2.45.

X	X^2		X	X^2
1	1		2	4
2	4		2	4
3	9		2	4
4	16		5	25
5	25		5	25
6	36		5	25
7	49		8	64
8	64		8	64
9	81		8	64
45	285		45	279

$$\sigma = \sqrt{\frac{285}{9} - \left(\frac{45}{9}\right)^2} = 2.58. \qquad \sigma = \sqrt{\frac{279}{9} - \left(\frac{45}{9}\right)^2} = 2.45.$$

If a distribution is so flat that the mid-values of each class closely approximate the class means, the value of σ (and π_2 and π_4) based on those mid-values may have a downward bias. Such a situation is unusual.

Sheppard's corrections may be applied when we are dealing with a continuous variable and high contact is present at both ends of the series. By "high contact" we mean that both tails of the distribution approach the X-axis asymptotically. If these conditions do not obtain, Sheppard's

[14] For a development, see H. L. Rietz (editor), *Handbook of Mathematical Statistics* pp. 92–94, Houghton Mifflin Company, Boston, 1924.

corrections should not be used, as the corrections may over-correct.[15] Neither is there justification for applying Sheppard's corrections if the original observations have not been made with reasonable accuracy.

In Table 48 the value of σ was found to be 4.51. If σ is computed for the ungrouped data of Table 25 (p. 165) the value obtained is 4.45. In Table 48 the value of π_2 (in terms of groups) is seen to be 5.0818. Applying the correction, we have

$$\mu_2 = 5.0818 - .0833 = 4.9975.$$

Computing the corrected value of σ gives

$$\sigma = 2.0\sqrt{4.9975} = 4.47,$$

which agrees somewhat more closely with the value of $\sigma = 4.45$, as obtained from the ungrouped data.

The α's and β's may be computed from the μ's in exactly the same way as from the π's. Thus

$$\alpha_1 = \frac{\mu_1}{\sigma} = \frac{\mu_1}{\sqrt{\mu_2}} = 0.$$

$$\alpha_2 = \frac{\mu_2}{\sigma^2} = \frac{\mu_2}{\sqrt{\mu_2^2}} = 1.$$

$$\sqrt{\beta_1} = \alpha_3 = \frac{\mu_3}{\sigma^3} = \frac{\mu_3}{\sqrt{\mu_2^3}}.$$

$$\beta_2 = \alpha_4 = \frac{\mu_4}{\sigma^4} = \frac{\mu_4}{\sqrt{\mu_2^4}} \text{ or } \frac{\mu_4}{\mu_2^2}.$$

Selected References

W. D. Baten: *Elementary Mathematical Statistics*, Chapters 3, 5; John Wiley and Sons, New York, 1938.

R. E. Chaddock. *Principles and Methods of Statistics*, Chapter IX; Houghton Mifflin Co., Boston, 1925.

F. E. Croxden and D. J. Cowden· *Practical Business Statistics*, Chapter X; Prentice-Hall Inc., 1934.

W. L. Crum, A. C. Patton, and A. R. Tebbutt: *Introduction to Economic Statistics*, Chapters XII, XIV; McGraw-Hill Book Co., New York, 1938.

F. C. Mills: *Statistical Methods Applied to Economics and Business* (Revised Edition), Chapter V and pages 448–450; Henry Holt and Co., New York, 1938.

L. H. C. Tippett: *The Methods of Statistics* (Second Edition), Chapter I; Williams and Norgate, London, 1937.

A. E. Waugh: *Elements of Statistical Method*, Chapter V and pages 113–128; McGraw-Hill Book Co., New York, 1938.

G. U. Yule and M. G. Kendall: *An Introduction to the Theory of Statistics* (Eleventh Edition), Chapters 8 and 9; Charles Griffin and Co., Ltd., London, 1937.

[15] See footnote 21 in Chapter XI (p. 301). Consult G. R. Davies and W. F. Crowder, *Methods of Statistical Analysis in the Social Sciences*, pp. 81–82, John Wiley and Sons, New York, 1933; also W. A. Shewhart, *Economic Control of Quality of Manufactured Product*, pp. 78–79, D. Van Nostrand Co., New York. 1931.

CHAPTER XI

DESCRIBING A FREQUENCY DISTRIBUTION BY A FITTED CURVE

A frequency distribution usually represents a sample drawn from a much larger population or universe. Even though a sample is composed of but a few hundred or a few score items, it may be reasonably representative of the larger universe from which it was drawn. Since it is virtually never possible to measure all of the individuals or items comprising a universe, we must form our notion of the larger group from a study of a sample. We may therefore fit any one of a number of types of curves to a fre-· quency distribution in order to attempt to describe what appears to be the general form of the curve for the entire population.

The purpose in fitting a curve to a frequency distribution may be any one of the following:

(1) To ascertain whether or not a given curve describes the general shape of the distribution. For example, we may wish to demonstrate that the chance errors involved in making some measurement may be described by a normal curve (see Chart 113).

(2) To enable us to generalize concerning the proportions of items which *should be expected* to fall above, below, or between certain values. For example, we may take the case of fitting a curve to a frequency distribution of the length of life of incandescent lamp bulbs; from such a procedure we are enabled to infer what proportion might, in general, be expected to burn 1,500 hours or more (or more or less than any specified number of hours). Similarly, in the case of the data shown in Charts 116, 117, and 119, we may determine the number of individuals which in general would be expected to occur above, below, or between any two X values. In like fashion the life insurance actuary may fit a curve to, or graduate data having to do with, deaths classified by age and thus de-termine the expected number of individuals dying during each year of life or surviving given ages.

(3) To enable us to determine, from a curve fitted to a given distribution, the probable distribution of values in a closely associated series. For

example, consider the illustration (developed elsewhere by the writers[1]) of a normal curve fitted to a distribution of the circumference of boys' heads; the curve enables us to draw a reasonable conclusion as to the number of caps of specified sizes which should be manufactured for the group from which the sample was drawn. Likewise, from a consideration of the

NUMBER
OF ITEMS

X VALUES
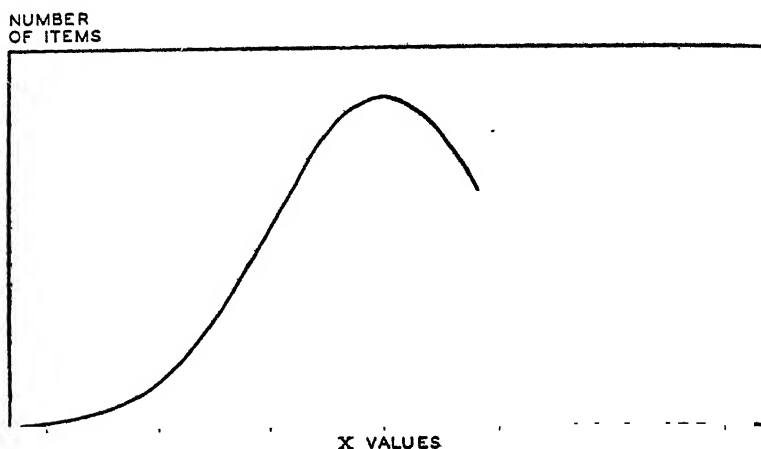
Chart 112. The Normal Curve of Error $Y_c = \dfrac{Ni}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}}$.

measurements of the circumference of men's necks (Chart 120), we can ascertain the probable number of collars of each size which would be needed.

This chapter will not attempt a comprehensive treatment of the topic of fitting frequency curves. For such a discussion, the reader is referred to the publications listed at the close of the chapter. We shall consider first the symmetrical curve known as the *normal curve of error* and then, briefly, binomials and certain of the simpler skewed curves.

The Normal Curve of Error

Development of the normal curve. The concept of the normal curve (pictured in Chart 112) appears to have been originally developed by Abraham De Moivre and explained in 1733 in a mathematical treatise[2]

[1] See F. E. Croxton and D J. Cowden, *Practical Business Statistics*. pp. 257–260, Prentice-Hall, Inc., New York, 1934

[2] *Approximatio ad Summam Terminorum Binomii* $(a + b)^n$ *in Seriem expansi*, Nov. 12, 1733, being a second supplement to *Miscellanea Analytica* 1730. See Karl Pearson, *Historical Note on the Origin of the Normal Curve of Errors*, Biometrika, Vol. 16 (1924), pp. 402–404; also, Helen M Walker, *Studies in the History of Statistical Method*, pp 13–17, 22–23. Williams and Wilkins, Baltimore, 1929

which its author believed had no practical applications other than as a solution of problems encountered in games of chance. Gauss later used the curve to describe the theory of accidental errors of measurements involved in the calculation of orbits of heavenly bodies. Because of Gauss' work this curve is sometimes referred to as the *Gaussian curve*.

Chart 113 shows a column diagram of 144 measurements of a line[3] and
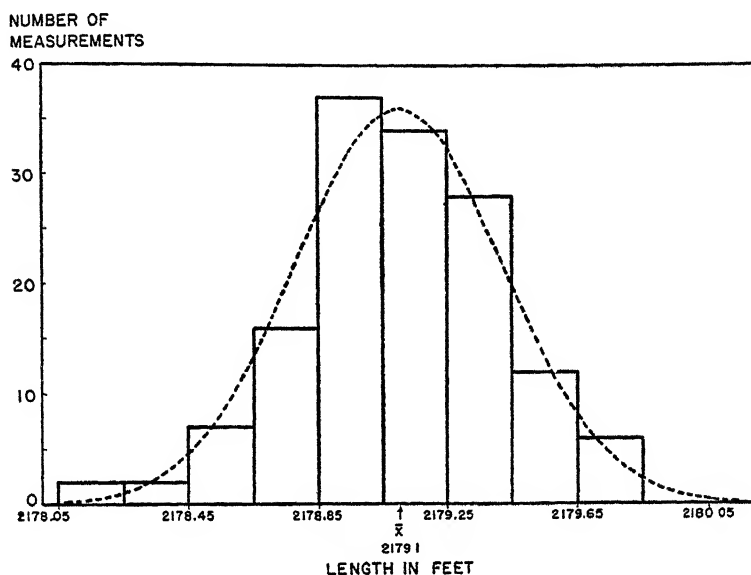
Chart 113. Normal Curve Fitted to 144 Measurements of the Length of a Line. (Measurements from L. D. Weld, *Theory of Errors and Least Squares*, p. 147, The Macmillan Company, New York, 1916.)

a normal curve of error fitted to these measurements. In fitting a normal curve it is assumed that only chance errors are present and that the arithmetic mean of the 144 measurements (2,179.1 feet) represents the best approximation of the true length of the line. It will be observed: (1) that small errors are more frequent than large ones, (2) that very large errors are unlikely to occur, and (3) that positive and negative errors of the same numerical magnitude are equally likely to occur—in other words, the curve is symmetrical. Because the fitted curve represents the relationship between the magnitude of an error and the probability of its occurrence in a given series of measurements, it is frequently termed the

---

[3] The 144 measurements are from L. D. Weld, *Theory of Errors and Least Squares*, p. 147. Macmillan. New York, 1916.

normal probability curve or simply the *normal* curve.[4] It will be seen in Chart 113 that the observed data have been shown as a rectangular frequency polygon, while the fitted data are represented by a curve drawn
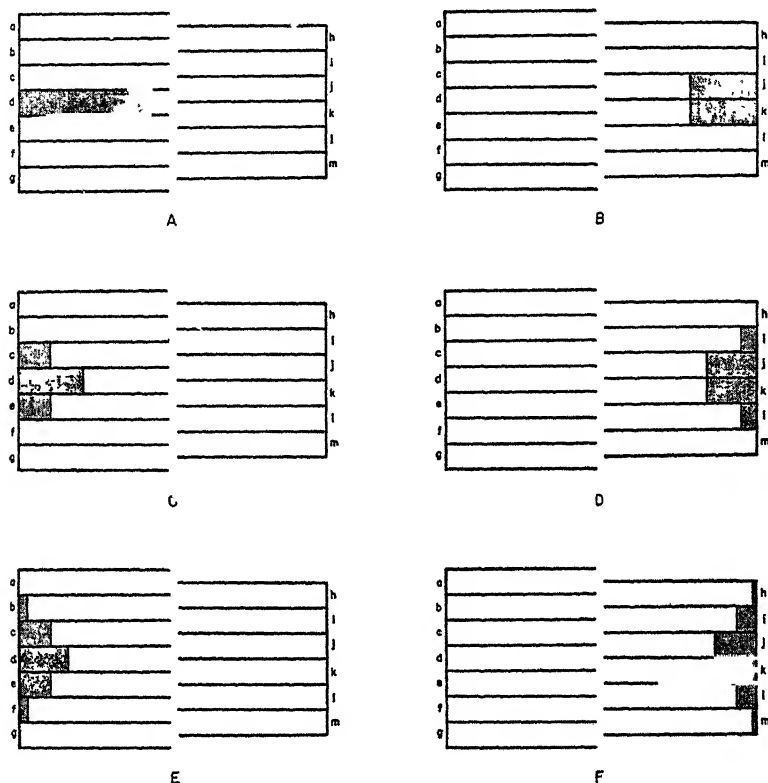
Chart 114.   Apparatus to Illustrate the Expansion of the Binomial $(\frac{1}{2} + \frac{1}{2})$.

with a dotted line.   Alternately, the observed data may be shown by a curve drawn with a solid line.

Chart 114 shows a simple apparatus which illustrates the play of chance in producing a symmetrical distribution.   The device consists of a number of troughs, open at one end and placed as shown in section A of Chart 114.

---

[4] See H. M. Goodwin, *Precision of Measurement and Graphical Measures*, p. 14, G. H. Ellis Co. (printers), Boston, 1909; also, A. DeF. Palmer, *Theory of Measurements*, p. 33, McGraw-Hill, New York, 1912.

The reader may be interested in consulting one authority who denies the applicability of the Gaussian curve as a description of errors of measurement.   See N. R. Campbell, *An Account of the Principles of Measurement and Calculation*, Ch. IX, especially p 182, note 1, Longmans, Green & Co., London, 1928.

Trough d is filled with sand or some similar granular substance. If the apparatus is tipped so that the left-hand side rises (section B of Chart 114), the sand in trough d will flow $\frac{1}{2}$ into trough j and $\frac{1}{2}$ into trough k. This represents the binomial $(\frac{1}{2} + \frac{1}{2})$.

If the right-hand side of the machine is then raised (section C of Chart 114), the sand from j will flow $\frac{1}{2}$ into c and $\frac{1}{2}$ into d, while the sand from k will flow $\frac{1}{2}$ into d and $\frac{1}{2}$ into e. Of the total amount of sand, we now have $\frac{1}{4}$ in c, $\frac{1}{2}$ in d, and $\frac{1}{4}$ in e, representing the expansion of the binomial $(\frac{1}{2} + \frac{1}{2})^2$. Again tipping the device, as in section D of Chart 114, $\frac{1}{2}$ of the sand from c flows into i, and $\frac{1}{2}$ into j; $\frac{1}{2}$ of the sand from d flows into j, and $\frac{1}{2}$ into k; and $\frac{1}{2}$ of the sand from e flows into k, and $\frac{1}{2}$ into l. The result is that $\frac{1}{8}$ of all the sand is in i, $\frac{3}{8}$ is in j, $\frac{3}{8}$ is in k, and $\frac{1}{8}$ is in l, representing the expansion of the binomial $(\frac{1}{2} + \frac{1}{2})^3$. Tipping the apparatus as in section E of Chart 114 causes the sand to flow $\frac{1}{16}$ into b, $\frac{4}{16}$ into c, $\frac{6}{16}$ into d, $\frac{4}{16}$ into e, and $\frac{1}{16}$ into f, representing the expansion of $(\frac{1}{2} + \frac{1}{2})^4$. Once more tipping the machine (section F of Chart 114) results in putting $\frac{1}{32}$ of the sand into h, $\frac{5}{32}$ into i, $\frac{10}{32}$ into j, $\frac{10}{32}$ into k, $\frac{5}{32}$ into l, and $\frac{1}{32}$ into m, which is the expansion of $(\frac{1}{2} + \frac{1}{2})^5$.

While the above illustration is instructive and gives us a picture of the expanded binomial, the device would become clumsy if we attempted to carry the expansion of the binomial much farther.[5] We may obtain similar results by tossing coins—a procedure which eliminates the necessity
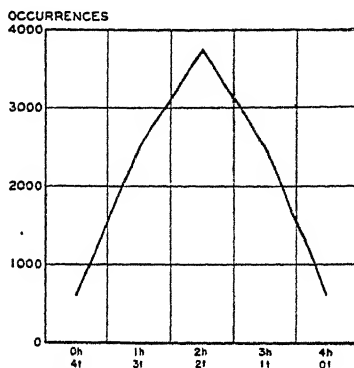
$$\frac{1}{16}t^4 + \frac{4}{16}ht^3 + \frac{6}{16}h^2t^2 + \frac{4}{16}h^3t + \frac{1}{16}h^4$$

Chart 115A. Expected Results of 10,000 Tosses of Four Coins.
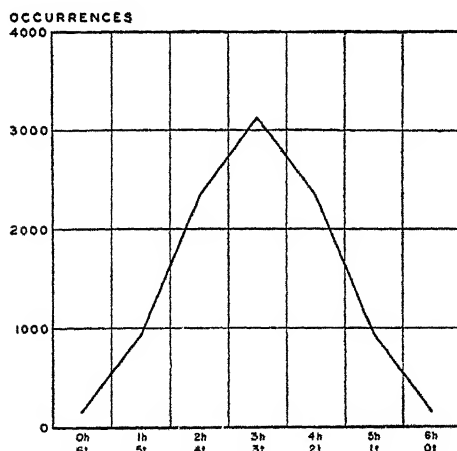
$$\frac{1}{64}t^6 + \frac{6}{64}ht^5 + \frac{15}{64}h^2t^4 + \frac{20}{64}h^3t^3 + \frac{15}{64}h^4t^2 + \frac{6}{64}h^5t + \frac{1}{64}h^6$$

Chart 115B. Expected Results of 10,000 Tosses of Six Coins.

[5] A slightly different device is shown in F. E Croxton and D. J. Cowden, *Practical Business Statistics*, p. 242, Prentice-Hall, Inc., New York. 1934.

of constructing any apparatus. It is assumed that we are tossing perfect coins which are evenly balanced and which will not stand on edge. With such a coin the chances of throwing a tail or a head are identical and may be expressed by $\frac{1}{2}t + \frac{1}{2}h$, where unity (1.0) represents certainty.

If two coins are tossed simultaneously, we may obtain either no heads (two tails), a tail and a head, or two heads. In order for two tails to appear, both coins must fall tails up. To obtain a tail and a head, one coin may show a tail and the other a head, or the first coin may show a head, the other a tail. Two heads may appear only if both coins show heads. Since a tail and a head may occur in two ways, while two tails may occur in but one way, it follows that there is twice as great a probability of throwing a tail and a head as of throwing two tails. Similarly, there is twice as great a chance of throwing a tail and a head as there is of throwing two heads.

$$\frac{1}{1024}t^{10} + \frac{10}{1024}ht^9 + \frac{45}{1024}h^2t^8 + \frac{120}{1024}h^3t^7 + \frac{210}{1024}h^4t^6$$
$$+ \frac{252}{1024}h^5t^5 + \frac{210}{1024}h^6t^4 + \frac{120}{1024}h^7t^3 + \frac{45}{1024}h^8t^2 + \frac{10}{1024}h^9t$$
$$+ \frac{1}{1024}h^{10}$$

Chart 115C. Expected Results of 10,000 Tosses of Ten Coins. (The probability of each combination is indicated by the binomial expansion shown under each part of Chart 115.)

We may express the probabilities arising from tossing two coins by

$$(\tfrac{1}{2}t + \tfrac{1}{2}h)^2,$$

in which the exponent 2 indicates the number of coins being tossed. Expanding this binomial gives

$$\tfrac{1}{4}t^2 + \tfrac{1}{2}th + \tfrac{1}{4}h^2.$$

Therefore, if two perfect coins are thrown 1,200 times, we could expect to obtain t^2 (no heads) 300 times, th (a head and a tail) 600 times, and h^2 (two heads) 300 times.

If three coins are tossed, we have the expression

$$(\tfrac{1}{2}t + \tfrac{1}{2}h)^3 = \tfrac{1}{8}t^3 + \tfrac{3}{8}t^2h + \tfrac{3}{8}th^2 + \tfrac{1}{8}h^3,$$

indicating that, if 1,200 throws were made, there should be no heads 150 times, one head and two tails 450 times, two heads and one tail 450 times, and three heads 150 times.

The results to be expected from tossing 4 coins is shown in section A of Chart 115, while the results to be expected from tossing 6 and 10 coins are shown respectively in parts B and C.   All of these curves are symmetrical and, as the number of coins tossed becomes greater, the curve becomes smoother.   When ten coins are tossed, there are eleven points to be plotted (See part C), but if 100 coins were tossed, there would be 101 points to plot and the curve would appear virtually the same as that of Chart 112. In fact, it can be shown that, as m becomes infinitely large, $(\frac{1}{2}t + \frac{1}{2}h)^m$ approaches as a limit the normal curve of error,[6] which is

$$Y_C = \frac{Ni}{\sigma\sqrt{2\pi}}\, e^{\frac{-x^2}{2\sigma^2}}.$$

The symbols are as follows:

$\quad Y_C$ = the computed height of an ordinate at the distance x from the arithmetic mean.

$\quad N$ = the number of observations in the sample.

$\quad i$ = the class interval.

$\quad \sigma$ = the standard deviation of the sample distribution.[7]

$\quad \pi$ = the constant, 3.14159; $\sqrt{2\pi}$ = 2.5066.

$\quad e$ = the constant, 2.71828, the base of the Naperian system of logarithms.

$\quad x$ = a selected deviation from the arithmetic mean.

Substituting the two constants mentioned above, we may write the equation

$$Y_C = \frac{Ni}{2.5066\sigma}\, 2.71828^{\frac{-x^2}{2\sigma^2}}.$$

Fitting the Normal Curve

In Chart 113 a normal curve was shown fitted to a series of measurements of a line.   It will be observed that those figures were repeated measurements of the same thing.   In Chart 116 we have a different type of data,

---

[6] See G. Udny Yule and M. G. Kendall, *An Introduction to the Theory of Statistics*. pp. 177–178, Charles Griffin and Company, London, 1937 (11th Edition); also D Caradog Jones, *A First Course in Statistics*, pp. 180–184, G. Bell and Sons, London, 1921. The exponent of the binomial is usually denoted by n.   We use m, however, since the symbol n will be used throughout later parts of the book to refer to "degrees of freedom."

[7] Strictly speaking, this should be the standard deviation of the universe, which we do not know. We may make an estimate of this value ($\bar{\sigma}$) from $\sqrt{\dfrac{\Sigma x^2}{N-1}}$, discussed in the following chapter   However, the difference between σ and $\bar{\sigma}$ is negligible when N is large, as in the case of the illustrations in this chapter.

representing measurements of a number of individuals from a homogene-
ous population. The chance errors involved in repeated measurements
of the same thing can be expected to follow a normal curve. However,
the measurements of a number of different individuals in respect to some
characteristic do not *necessarily* follow such a curve. A distribution of
the heights of a homogeneous group of adult individuals, for example,
could be expected to be essentially normal, but a distribution of the weights
of the same individuals would be noticeably skewed to the right. While
the basal diameter of the egg-capsules of the snails in Chart 116 is well
described by the fitted normal curve, it is quite likely that the weights of
these same eggs would show definite skewness.
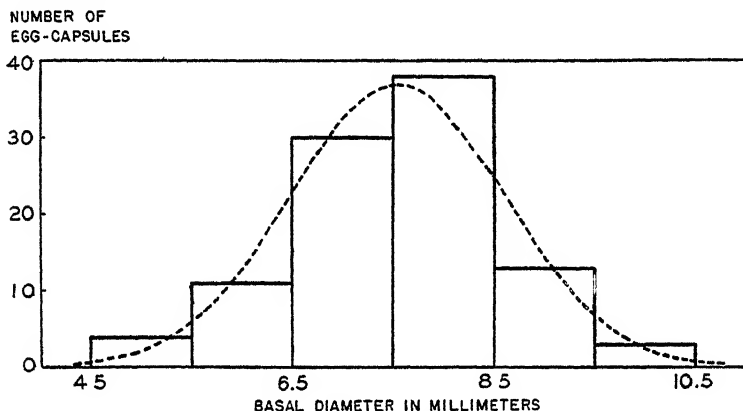
Chart 116. Normal Curve Fitted to Basal Diameters of 99 Egg-Capsules of a Marine
Snail, *Sipho curtus.* (Data of basal diameters from Gunnar Thorson, *Studies on the
Egg-Capsules and Development of Arctic Marine Prosobranchs*, p 7, Meddelelser om
Grønland -udgione af- Kommissionen for Videnskabelige Endersøgelser i Grønland.)

The fitted curve in Chart 116 indicates the shape of the distribution
we should expect if our sample were much larger, or if we had measured
the entire population. It implies that, if a larger group were studied, we
should find a few instances with basal diameters both smaller and larger
than those found in the sample.

Fitting the normal curve to data of physical ability. The data of Table
55 show a distribution of the distances which 303 high school freshman
girls were able to throw a baseball. It may be observed that very few
of the girls threw the baseball less than 45 feet and very few threw it 115
feet or farther. The column diagram of this distribution is shown in
Chart 117. The distribution tends to be symmetrical, and we infer that a
normal curve might reasonably be fitted. We shall, first, determine the
values of a number of ordinates in order to ascertain the exact outline of

the fitted curve and, second, compute the theoretical frequencies to be expected in each class of the distribution.

TABLE 55

BASEBALL THROWS FOR DISTANCE BY 303 FIRST-
YEAR HIGH SCHOOL GIRLS

Distance in feet	Number of girls
15 but under 25	1
25 but under 35	2
35 but under 45	7
45 but under 55	25
55 but under 65	33
65 but under 75	53
75 but under 85	64
85 but under 95	44
95 but under 105	31
105 but under 115	27
115 but under 125	11
125 but under 135	4
135 but under 145	1
Total	303

Source· Leonora W Stewart and Helen West, The Froebel School, Gary, Indiana Measurements were made in 1935.
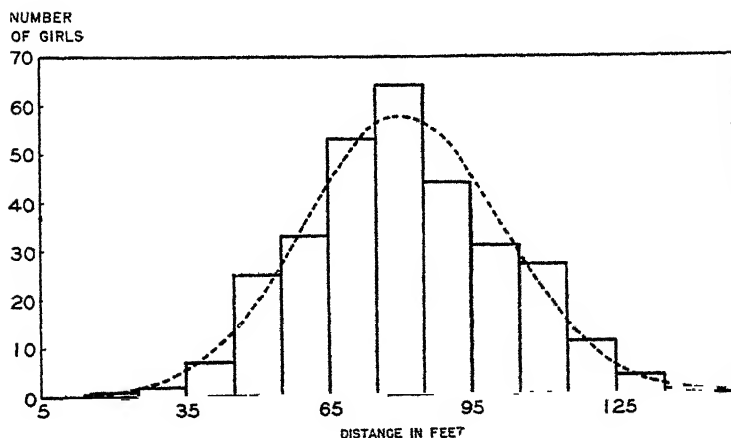
Chart 117.   Normal Curve Fitted to Data of Baseball Throws for Distance by First Year High School Girls.   (Data from Tables 55 and 56.)

Referring again to the formula for the normal curve

$$Y_C = \frac{Ni}{2.5066\sigma}\, 2.71828^{\frac{-x^2}{2\sigma^2}},$$

it appears that we need the values of N, $\overline{X}$, and σ in order to fit a normal curve to a distribution. Computing by procedures described in preceding chapters, we find $\overline{X} = 80.63$ feet and $\sigma = 20.95$ feet. (For comparison $\tilde{\sigma} = 20.98$ feet.) As there were 303 girls, $N = 303$.

We shall first compute the ordinate to be erected at the mean. This is designated as Y_o and is the maximum ordinate of the fitted curve. Since $x = 0$ at the mean, we have

$$Y_o = \frac{303 \times 10}{2.5066 \times 20.95}\, 2.71828^{\frac{-0^2}{2(20\,95)^2}}.$$

In the expression above, the exponent of 2.71828 is zero. Since a number raised to the zero power is 1, $2.71828^{\frac{-0^2}{2(20\,95)^2}} = 1$. It is apparent then that the expression $e^{\frac{-x^2}{2\sigma^2}}$ is always equal to 1 for the ordinate erected at the mean (Y_o) and

$$Y_o = \frac{Ni}{\sigma\sqrt{2\pi}}.$$

Therefore

$$Y_C = \frac{Ni}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}} = Y_o\, 2.71828^{\frac{-x^2}{2\sigma^2}}.$$

For the problem in hand

$$Y_o = \frac{303 \times 10}{2.5066 \times 20.95} = 57.7.$$

We now wish to erect enough additional ordinates on either side of Y_o to enable us to sketch a reasonably smooth curve. If we select successive distances of 4.19 feet from the mean, we shall erect ordinates at steps of $\frac{1}{5}\sigma$ from the mean. The first pair of ordinates (since the curve is symmetrical) are to be erected at $x = \pm 4.19$ feet from the mean ($X = 84.82$ and 76.44 feet), using the expression

$$Y_C = 57.7 \times 2.71828^{\frac{-(4\,19)^2}{2(20.95)^2}}.$$

In order to determine the value Y_C, it is not necessary to compute $2.71828^{\frac{-(4\,19)^2}{2(20.95)^2}}$ but merely to refer to Appendix D. Looking up the ap-

propriate value of $\frac{x}{\sigma}$, which in this case is $\frac{4.19}{20.95} = .20$, we find that

$$2.71828^{\frac{-(4\ 19)^2}{2(20\ 95)^2}} = .98020$$

and

$$Y_C = 57.7 \times .98020 = 56.6.$$

For the next pair of ordinates, $x = \pm 8.38$ feet ($X = 89.01$ feet and 72.25 feet) and

$$Y_C = 57.7 \times 2.71828^{\frac{-(8\ 38)^2}{2(20\ 95)^2}}.$$

Here the ratio of $\frac{x}{\sigma}$ is .40 and, referring to Appendix D, we have

$$Y_C = 57.7 \times .92312 = 53.3.$$

The process of determining the heights of the ordinates can be handled most expeditiously by use of a table similar to Table 56. The ordinates in the upper and lower parts of the table are identical since the fitted curve is symmetrical.

The fitted curve is shown in Chart 117. It follows the general shape of the sample, but smooths out the irregularities and indicates what might be expected if the performance of a very large number of comparable girls could be recorded. What we have done so far gives merely the shape of the fitted curve and a visual impression of the suitability of the fit, which appears good in this instance.

We have not yet undertaken to say what proportion of girls should be expected to throw a baseball 100 feet or more, or what proportion might be expected to throw a baseball 50 feet or less. Neither have we attempted to say what proportion of girls should theoretically be expected to fall into the various classes. The expected frequencies in each class are ascertained by integrating the fitted curve.[8] However, the procedure is greatly simplified by making use of a table of the areas under the normal curve (Appendix E). Referring to Table 57, we determine the expected frequencies in each class as follows:

1. In column (1) of the table, enter the classes of the original distribution, allowing for one or two additional classes at each end, since the fitted curve should usually have a greater range than the sample. Theoretically the fitted curve is of unlimited range in both directions. Allow two spaces for the class in which the mean falls.

---

[8] When there is a fairly large number of classes in the sample distribution, the theoretical frequencies may be ascertained with reasonable accuracy by erecting ordinates at the mid-value of each class See F. E. Croxton and D. J Cowden, *Practical Business Statistics* pp 249–251, Prentice-Hall, Inc, New York, 1934.

TABLE 56

DETERMINATION OF ORDINATES OF NORMAL CURVE FITTED TO DATA OF BASEBALL
THROWS FOR DISTANCE BY FIRST-YEAR HIGH SCHOOL GIRLS

($\overline{X} = 80\ 63$ feet; $\sigma = 20\ 95$ feet; $Y_o = 57\ 7$)

X (in feet, where ordinates are to be erected)	x (in feet, deviation of X from $\overline{X}$)	$\dfrac{x}{\sigma}$	Proportionate height of ordinate $2\ 71828^{\frac{-x^2}{2\sigma^2}}$ (Appendix D)	Height of ordinate [Col. 4 $\times$ Y_o]
(1)	(2)	(3)	(4)	(5)
13 59	$-67\ 04$	3 20	00598	3
17.78	$-62\ 85$	3 00	01111	6
21 97	$-58\ 66$	2 80	.01984	1.1
26 16	-54.47	2 60	03405	2.0
30.35	-50.28	2 40	.05614	3 2
34 54	$-46\ 09$	2.20	.08892	5.1
38 73	$-41\ 90$	2.00	13534	7.8
42 92	$-37\ 71$	1 80	.19790	11.4
47.11	$-33\ 52$	1.60	.27804	16.0
51.30	-29.33	1 40	.37531	21 7
55.49	-25.14	1.20	.48675	28 1
59 68	-20.95	1 00	.60653	35.0
63.87	$-16\ 76$	.80	.72615	41.9
68.06	$-12\ 57$	.60	.83527	48 2
72.25	$-\ 8.38$	.40	.92312	53.3
76.44	$-\ 4.19$	.20	98020	56 6
80.63	0	0	1 00000	57 7
84.82	$+\ 4.19$	20	98020	56 6
89 01	$+\ 8 38$	.40	92312	53.3
93 20	$+12\ 57$	.60	.83527	48 2
97 39	$+16\ 76$	.80	72615	41.9
101 58	$+20.95$	1.00	.60653	35.0
105.77	$+25.14$	1.20	.48675	28.1
109 96	$+29.33$	1.40	.37531	21.7
114.15	$+33.52$	1 60	27804	16 0
118.34	$+37.71$	1 80	.19790	11.4
122 53	$+41\ 90$	2 00	.13534	7.8
126 72	$+46.09$	2.20	.08892	5.1
130.91	$+50\ 28$	2.40	.05614	3.2
135 10	$+54.47$	2.60	.03405	2.0
139.29	$+58.66$	2.80	.01984	1.1
143.48	$+62\ 85$	3.00	.01111	.6
147 67	$+67\ 04$	3.20	.00598	.3

TABLE 57

DETERMINATION OF EXPECTED FREQUENCIES IN EACH CLASS FOR BASEBALL THROWS FOR DISTANCE BY FIRST-YEAR HIGH SCHOOL GIRLS

($\bar{X}$ = 80 63 feet; σ = 20.95 feet)

Distance in feet (1)	Limits of classes — Lower limits (2)	Limits of classes — Upper limits (3)	x deviation from mean to limit (4)	$\dfrac{x}{\sigma}$ (5)	Per cent of area between mean and limit (Appendix E) (6)	Per cent of area in each class (7)	Expected frequencies in each class N = 303* (8)
Under 5				..	50.00	01	.2
5 but under 15	5		75 63	3.61	49 99	08	9
15 but under 25	15		65 63	3 13	49 91	30	3 2
25 but under 35	25		55 63	2 66	49 61	1 07	9.1
35 but under 45	35		45 63	2.18	48 54	3 00	20 2
45 but under 55	45		35 63	1 70	45 54	6 66	35 0
55 but under 65	55		25 63	1 22	38 88	11 54	50.6
65 but under 75	65		15 63	.75	27 34	16 70	57 4
75 but under 85	75	85	5.63 / 4.37	27 / 21	10 64 / 8 32	18 96	52 0
85 but under 95		95	14 37	69	25 49	17 17	37 0
95 but under 105		105	24 37	1.16	37 70	12 21	22 0
105 but under 115		115	34 37	1 64	44 95	7 25	10 2
115 but under 125		125	44 37	2.12	48 30	3 35	3.7
125 but under 135		135	54 37	2.60	49 53	1.23	1 1
135 but under 145		145	64 37	3 07	49 89	36	3
145 but under 155		155	74.37	3 55	49 98	09	.1
155 and over				..	50 00	02	
Total	..	..		..		100 00	303 0

277

* One decimal is usually shown in this column in order that the total of the expected frequencies will agree, to within 1 or 2, with the total of the observed frequencies. This is of importance in making the χ² test of Table 61

2. In column (2), write the lower limits of each class below the mean in value and the lower limit of the class which contains the mean.

3. In column (3), write the upper limit of each class above the mean in value and the upper limit of the class which includes the mean.

4. The process of determining the expected frequencies in each class uses the mean as the basis of reference and involves, first, those classes above (or below) the mean in value and, then, those below (or above). We shall therefore ascertain first the expected frequencies between the mean (80.63 feet) and the upper limit (85 feet) of the class in which the mean falls. The deviation x of the upper limit from the mean is 4.37
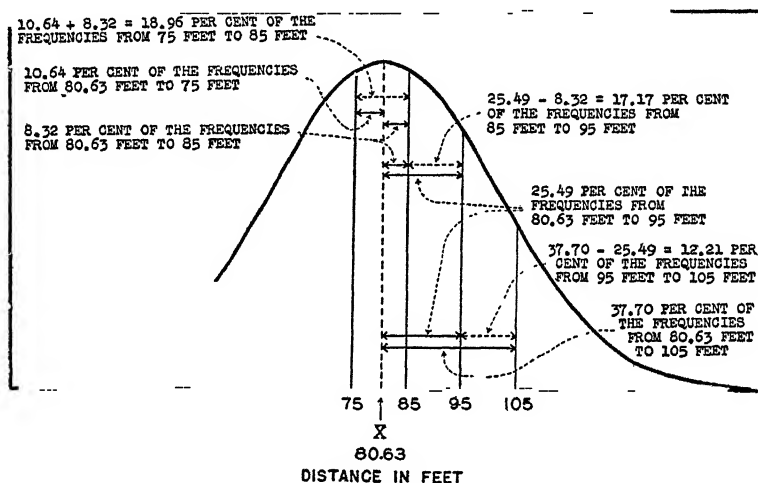
Chart 118. Graphic Representation of Procedure in Columns (6) and (7) of Table 57.

feet; this value is entered in column (4). Instead of integrating the curve from $X = 80.63$ feet to $X = 85$ feet, we determine the value of $\frac{x}{\sigma}$ and make use of Appendix E. Since $\sigma = 20.95$ feet,

$$\frac{x}{\sigma} = \frac{4.37}{20.95} = .21.$$

This value is entered in column (5). Now, looking up .21 in Appendix E, we find .0832, indicating that .0832 of the area of the normal curve (the total area being expressed as 1.00000) is between the mean and 85 feet. In other words, 8.32 per cent of the frequencies would be expected to fall between the mean and 85 feet. This value is entered in column (6). The procedure is shown graphically in Chart 118.

5. The next step consists of determining the expected frequencies be-

tween the mean and the upper limit of the first class above the mean. This limit is 95 feet; $x = 14.37$ feet and

$$\frac{x}{\sigma} = \frac{14.37}{20.95} = .69.$$

Looking up .69 in Appendix E shows that 25.49 per cent of the frequencies would be expected to occur between the mean and 95 feet.   This value is entered in column (6).   If 25.49 per cent of the items fall between 80.63 and 95 feet, while 8.32 per cent of the items fall between 80.63 and 85 feet, there would be $25.49 - 8.32 = 17.17$ per cent of the items between 85 and 95 feet.   The result of this subtraction is entered in column (7); this procedure is also indicated graphically in Chart 118.

6.   The procedure in step 5 is repeated for each class above the mean in value.   The expected frequencies from the mean to the upper limit of each class are ascertained, and then the frequencies from the mean to the upper limit of the preceding class are subtracted as shown in the table.

7.   The expected frequencies falling between the mean and the lower limits shown in column (2) of the table are next determined.   Since these areas are also cumulative, successive subtraction is again necessary.

8.   We now have entered in column (7) the expected frequencies for each class except the class containing the mean.   We have determined, in column (6), that there are 8.32 per cent of the expected frequencies from the mean to 85 feet and that there are 10.64 per cent of the expected frequencies from the mean to 75 feet.   Adding these two figures gives 18.96 per cent, the proportion of expected frequencies falling in this class [see column (7) and Chart 118].

9.   The total of column (7) should be 100.00, as there are 50.00 per cent of the expected frequencies from the mean to either extreme of the distribution.   In order to see the agreement between the observed and the expected frequencies, we include column (8), which is obtained by multiplying 303 by the expected frequency of each class and dividing by 100, or, in a single operation, by multiplying 3.03 by the expected frequency of each class.[9]

A comparison of the expected frequencies, shown in column (8) of Table 57, with the observed frequencies of Table 55 reveals a general agreement of the figures, the difference being greatest for the class "85 but under 95

[9] For the class 75 but under 85 feet,

$$18.96 : 100.00 :: f_c : 303$$
$$100.00 f_c = 303 \times 18\,96$$
$$f_c = \frac{303}{100.00}\,18\,96 = 3.03 \times 18.96 = 57\,4.$$

feet." A test of the "goodness of fit" of the normal curve, based upon the expected frequencies, will be described later.

We have not yet answered the question: "What proportion of girls should be expected to throw a baseball 100 feet or more?" The proportion of expected frequencies from the mean (80.63 feet) to 100 feet ($x = 19.37$ feet) is obtained by computing

$$\frac{x}{\sigma} = \frac{19.37}{20.95} = .92,$$

and referring to Appendix E. The proportion is .3212, or 32.12 per cent. Since 50 out of 100 girls would be expected to throw 80.63 feet or more, it

NUMBER OF
PLAYERS
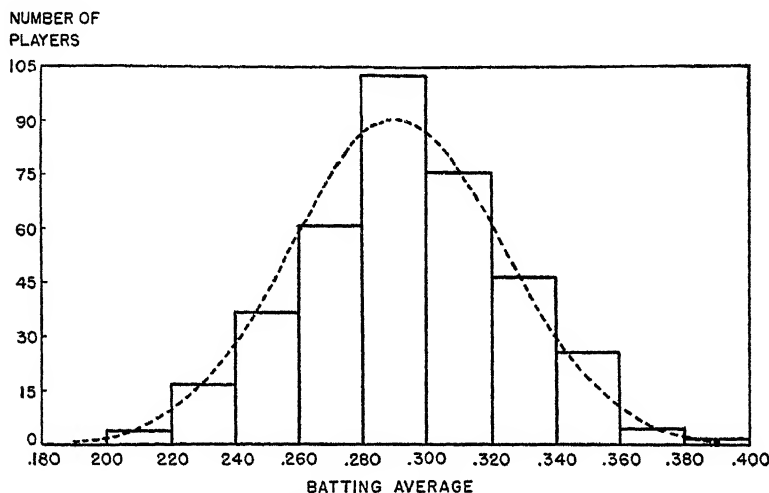
Chart 119. Normal Curve Fitted to Batting Averages of 379 Major and Minor League Baseball Players, 1936. (Included are only those regular players who were in 75 or more games and at bat 225 or more times. Source: David L. Rolbein.)

follows that 50.00 − 32.12 = 17.88, or 17.9 per cent, would in general throw 100 feet or more.

To determine the proportion expected to throw 50 feet or less, the procedure is similar.

$$x = 80.63 \text{ feet} - 50 \text{ feet} = 30.63 \text{ feet},$$
$$\frac{x}{\sigma} = \frac{30.63}{20.95} = 1.46.$$

Looking this up in Appendix E, we find .4279, or 42.79 per cent. Subtracting from 50.00 leaves 7.2 per cent as the proportion expected to throw 50 feet or less.

The normal curve fitted to batting averages. David L. Rolbein has supplied the following interesting illustration of a distribution of a human

ability which appears to be adequately described by means of a normal curve (see Chart 119). The observed data of batting averages included all players in the American League, the National League, the American Association, the International Association, and the Pacific Coast League who had played in 75 or more games and who had been at bat 225 or more times. The data therefore included only regular players, $N = 379$. The fitting constants were $\overline{X} = .2942$, while $\sigma = 1.67$ classes, or .0334 units (the class intervals were .020).

Granting unchanged conditions as to skill of batters, pitchers, and fielders and as to liveliness of the baseball, etc., what proportion of regular players

TABLE 58

NECK CIRCUMFERENCE OF 231
MALE COLLEGE STUDENTS

Mid-values (in inches)	Number of students
12.5	4
13 0	19
13 5	30
14.0	63
14 5	66
15 0	29
15 5	18
16 0	1
16 5	1
Total	231

Source· Confidential.

would be expected to bat .350 or better? Since $\overline{X} = .2942$, $x = .350 - .2942 = .0558$, and $\dfrac{x}{\sigma} = \dfrac{.0558}{.0334} = 1.67$. From Appendix E, $\dfrac{x}{\sigma} = 1.67$ gives .4525 (or 45.25 per cent) as the proportion batting between .2942 and .350. Therefore $50.00 - 45.25 = 4.75$ per cent would, in general, be expected to bat .350 or better.

The normal curve and collar sizes. To illustrate this use of the normal curve, let us assume that a maker of collars is considering the production of a collar styled especially for college men. Consideration will, of course, be given to the number of collars of each size which should be made. Since college men represent a selected group, it would be desirable to adjust the manufacturing schedule to their particular requirements. Extensive data on the circumference of the necks of college men are not available, but in Table 58 are shown the neck measurements of 231 male college

students.   To fit a normal curve, we need $\overline{X} = 14.232$ inches and $\sigma = .719$ inches.   The column diagram of the observed data and the fitted curve are shown in Chart 120.

Our problem, in this instance, is not to determine the expected proportion of college men having necks "12.75 but under 13.25" inches in circumference, "13.25 but under 13.75" inches in circumference, etc., but rather to determine the number of collars of each size (by half sizes) which should be made.   Experience shows that, on the average, collars are worn about $\frac{3}{4}$ of an inch larger than the circumference of the neck.   This means that collars size 14 would be worn by men whose necks averaged 13.25 inches and, since we are dealing with half sizes, the necks would range from 13 to
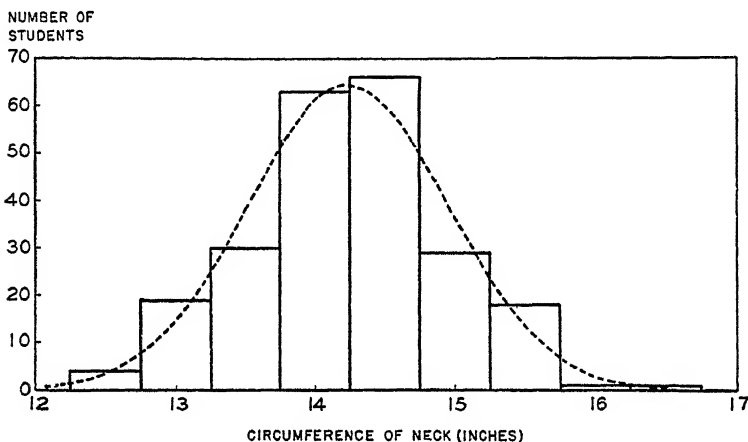
Chart 120.   Normal Curve Fitted to Neck Circumference of 231 Male College Students.
(Based on data of Table 58.)

13.5 inches in circumference.   The first column of Table 59 lists the collar sizes, while the second column shows the corresponding neck circumferences.   It is for these classes that we need to ascertain the theoretical frequencies.   This is done in the remainder of the columns and the expected frequencies ($N = 1,000$) are shown in column (9).   If our basic data are representative, there would be about 270 customers in a thousand calling for size 15 collars, 221 asking for size $14\frac{1}{2}$, 213 requesting size $15\frac{1}{2}$, etc.   It is interesting to observe that we might expect only 8 out of a thousand of this group to ask for size 13 or smaller and but 7 out of a thousand to require 17 or larger.

Suitability of the normal curve.   As previously pointed out, the normal curve is only one of a number of kinds of curves which may be fitted to a frequency distribution.   It should in no sense be thought of as a form having general applicability to all distributions.   Since this is true, what

TABLE 59

DETERMINATION OF COLLAR SIZES FOR MALE COLLEGE STUDENTS

($\overline{X}$ = 14.232 inches; σ = .719 inches)

Collar size (1)	Corresponding neck circumference (2)	Lower limits (3)	Upper limits (4)	x from mean to limit (5)	$\frac{x}{\sigma}$ (6)	Cumulative theoretical frequencies as percentages (Appendix E) (7)	Theoretical frequencies as percentages (8)	Theoretical frequencies $N = 1000$ (9)
...	Smaller than 11.5			2.732	3.80	50.00	.01	.1
12½	11.5 but under 12.0	11.5		2.232	3.10	49.99	.09	.9
13	12.0 but under 12.5	12.0		1.732	2.41	49.90	.70	7.0
13½	12.5 but under 13.0	12.5		1.232	1.71	49.20	3.56	35.6
14	13.0 but under 13.5	13.0		.732	1.02	45.64	11.03	110.3
14½	13.5 but under 14.0	13.5		.232	.32	34.61	22.06	220.6
15	14.0 but under 14.5	14.0	14.5	.268	.37	{ 12.55 / 14.43	26.98	269.8
15½	14.5 but under 15.0		15.0	.768	1.07	35.77	21.34	213.4
16	15.0 but under 15.5		15.5	1.268	1.76	46.08	10.31	103.1
16½	15.5 but under 16.0		16.0	1.768	2.46	49.31	3.23	32.3
17	16.0 but under 16.5		16.5	2.268	3.15	49.92	.61	6.1
17½	16.5 but under 17.0		17.0	2.763	3.85	49.99	.07	.7
...	17.0 or larger					50.00	.01	.1
Total		...	...				100.00	1,000.0

guides are there which will tell us when to fit a normal curve, or when fitted if it is suitable?

1. The plotted curve or column diagram of the sample distribution serves as a very crude guide. If there is skewness present, it will be apparent, as will also any irregularities.

2. The sample data may be cumulated and put into percentage form as in Table 60; these cumulative percentages may then be plotted on arithmetic probability paper[10] as in Chart 121. If the resulting curve is approximately a straight line, we may proceed with assurance to fit a normal curve.

3. The moments of the sample distribution may be computed as described in Chapter X. From these we may compute

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \text{ (a measure of skewness)},$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} \text{ (a measure of kurtosis)},$$

$$\kappa_2 = \frac{\beta_1(\beta_2 + 3)^2}{4(4\beta_2 - 3\beta_1)(2\beta_2 - 3\beta_1 - 6)} \text{ (a general}$$

measure of departure from normal)

TABLE 60

CUMULATIVE DISTRIBUTION OF BASEBALL THROWS
FOR DISTANCE BY 303 FIRST-YEAR HIGH
SCHOOL GIRLS

Distance in feet	Number of girls	Per cent of total
Less than 25	1	33
Less than 35	3	99
Less than 45	10	3.30
Less than 55	35	11.55
Less than 65	68	22 44
Less than 75	121	39.93
Less than 85	185	61 06
Less than 95	229	75.58
Less than 105	260	85.81
Less than 115	287	94.72
Less than 125	298	98.35
Less than 135	302	99.67
Less than 145	303	100.00

Source: Cumulative data of Table 55.

---

[10] The vertical scale is so designed that the ogive of a normal curve will appear as a straight line. Paper available from the Codex Book Co., Norwood, Mass.
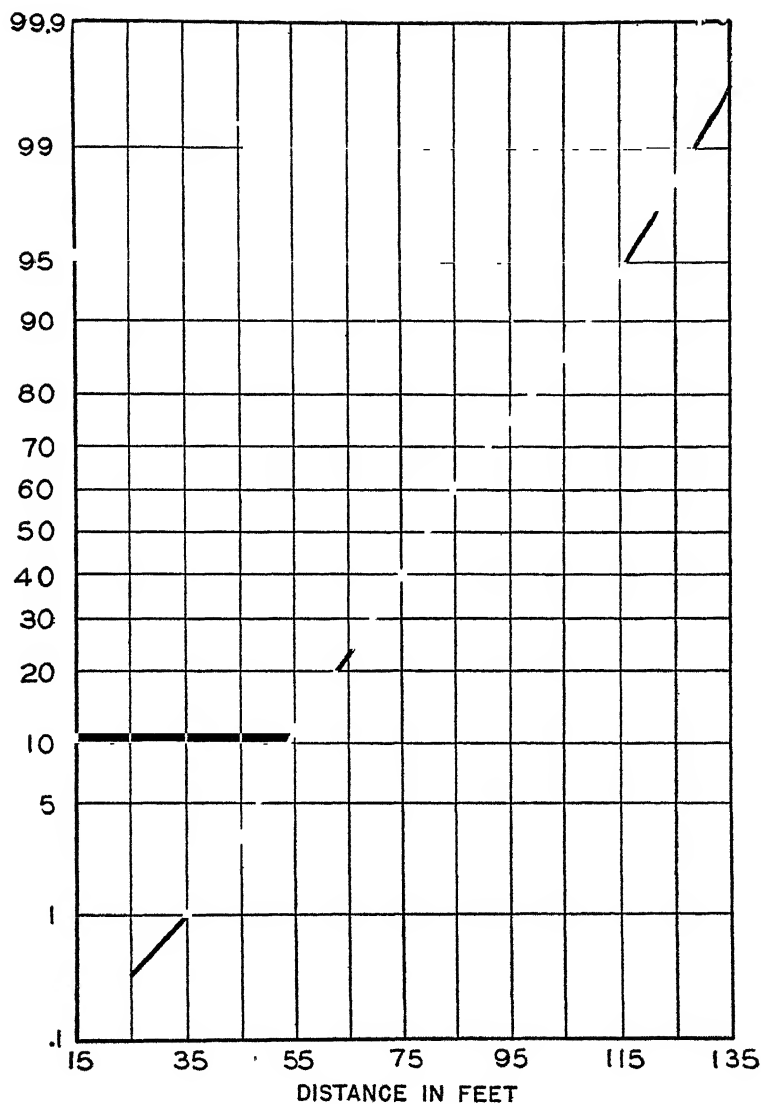
Chart 121. Baseball Throws for Distance by 303 First-Year High School Girls, Shown on Arithmetic Probability Paper. (Based on data of Table 60.)

If the distribution is essentially normal, we should obtain $\beta_1 = 0$, $\beta_2 = 3$, and $\kappa_2 = 0$. For the throws of a baseball by high school freshman girls, we find $\beta_1 = .010415$, $\beta_2 = 2.772352$, and $\kappa_2 = -016062$, indicating that a normal curve can probably be used to describe the series.[11] Certain other values of β_1, β_2, and κ_2 indicate the suitability of other curves of the Pearson group.[12]

4. After the curve has been fitted and the theoretical frequencies have been determined, the number of items in the sample are prorated among the classes upon this basis. For the baseball throws, $N = 303$, and the theoretical frequencies for a total of 303 are shown in the last column of Table 57. These theoretical frequencies and the observed frequencies are then compared by means of the χ^2 test.

$$\chi^2 = \Sigma \frac{(f - fc)^2}{fc},$$

where f is an observed frequency in a class and fc is the corresponding theoretical frequency. Table 61 shows the computation of χ^2 for the data of girls' baseball throws.

The goodness of fit is indicated by χ^2 when considered in conjunction with n, the number of degrees of freedom. The number of degrees of freedom is obtained by subtracting from the number of classes the number of degrees of freedom lost in the fitting process.[13] In this case 3 degrees of freedom were lost because the original data and the fitted data were made to agree in respect to the number of items (N), the mean ($\bar{X}$), and the standard deviation (σ). χ^2 and n enable P to be determined,[14] which tells us the probability that a fit as bad or worse might occur because of chance variations of sampling.

On account of the great effect upon χ^2 of differences between small observed and expected frequencies at the ends of a distribution, it is generally necessary to combine two or more classes at each end. Fisher suggests that no group should contain fewer than 5 expected frequencies. Combining the first three classes and the last two classes leaves 10 classes, as

---

[11] A somewhat more satisfactory test suggested by R A Fisher is given in Appendix B, section XI–1 For a discussion of the reliability of β_1 and β_2, see L H. C Tippett, *The Methods of Statistics*, p 86, Williams and Norgate, London, 1937 (2nd Edition).

[12] See Karl Pearson, *Tables for Statisticians and Biometricians*, pp. lx, f, University Press, Cambridge, 1914; and W. P. Elderton, *Frequency Curves and Correlation*, (3rd Edition) Chs. IV and V, Cambridge University Press, Cambridge, England, 1938.

[13] See Chapter XII, p. 312, for a more complete statement concerning degrees of freedom.

[14] See Appendix I for a table of P. The chi-square test is discussed in R. A. Fisher, *Statistical Methods for Research Workers*, Ch IV, Oliver and Boyd, Edinburgh, 1938 (7th edition) and in L. H. C. Tippett, *The Methods of Statistics*, Ch. IV, Williams and Norgate, London, 1937 (2nd Edition).

shown in Table 61; we have $\chi^2 = 6.39$, $n = 7$, and P is about .50. These results indicate that the normal curve is a good description of the series, since, if the distribution of distances thrown is actually normal, we might expect a fit as bad or worse than this about 50 times out of a hundred, because of chance variations attributable to sampling. A rule of thumb is often undesirable because inflexible, but we may regard a P of less than 05 as indicating a poor fit.

TABLE 61

CHI-SQUARE TEST OF GOODNESS OF FIT FOR NORMAL CURVE FITTED TO BASEBALL THROWS FOR DISTANCE BY FIRST-YEAR HIGH SCHOOL GIRLS

Distance in feet (1)	f observed frequency (2)	f_c expected frequency (3)	$f - f_c$ (4)	$(f - f_c)^2$ (5)	$\dfrac{(f - f_c)^2}{f_c}$ (6)
15 but under 25	1	1 1			
25 but under 35	2	3 2	} -3.4	11.56	.86
35 but under 45	7	9 1			
45 but under 55	25	20.2	4 8	23 04	1.14
55 but under 65	33	35.0	$-2 0$	4 00	.11
65 but under 75	53	50 6	2 4	5 76	.11
75 but under 85	64	57 4	6 6	43 56	.76
85 but under 95	44	52 0	$-8 0$	64.00	1.23
95 but under 105	31	37.0	-6.0	36.00	.97
105 but under 115	27	22 0	5.0	25 00	1.14
115 but under 125	11	10 2	8	.64	.06
125 but under 135	4	3.7	} $- .2$	.04	.01
135 but under 145	1	1.5			
Total . . .	303	303.0	0	. . .	6.39

$$\chi^2 = 6.39; \quad n = 10 - 3 = 7.$$

Binomials

It was previously shown that the expansion of a symmetrical binomial $(\frac{1}{2} + \frac{1}{2})^m$ can be approximated experimentally by tossing coins. An *asymmetrical* binomial may be expanded experimentally in a similar fashion.

Experimental construction of skewed binomials. Let us consider, first, a single die two sides of which are colored black. If we toss this die, it is apparent that the probability (p) of having a black side come up is 1 out of 3, or $\frac{1}{3}$, while the probability $(q = 1 - p)$ of obtaining a white side is 2 out of 3, or $\frac{2}{3}$. We may express the situation as $qw + pb$ or $\frac{2}{3}w + \frac{1}{3}b$, which indicates that, if the die (assumed to be perfectly balanced) is tossed 1,500 times, we should expect a white side to appear 1,000 times and a black side 500 times.

If, now, we toss two dice (each having two black sides), there may appear either no black faces (2 white faces), a black face and a white face, or 2 black faces. The expression is

$$(\tfrac{2}{3}w + \tfrac{1}{3}b)^2 = \tfrac{4}{9}w^2 + \tfrac{4}{9}wb + \tfrac{1}{9}b^2.$$

Therefore, if 1,800 throws are made, we should expect to obtain no black faces (two white faces) 800 times, a black face and a white face 800 times, and two black faces 200 times.

OCCURRENCES IN THOUSANDS
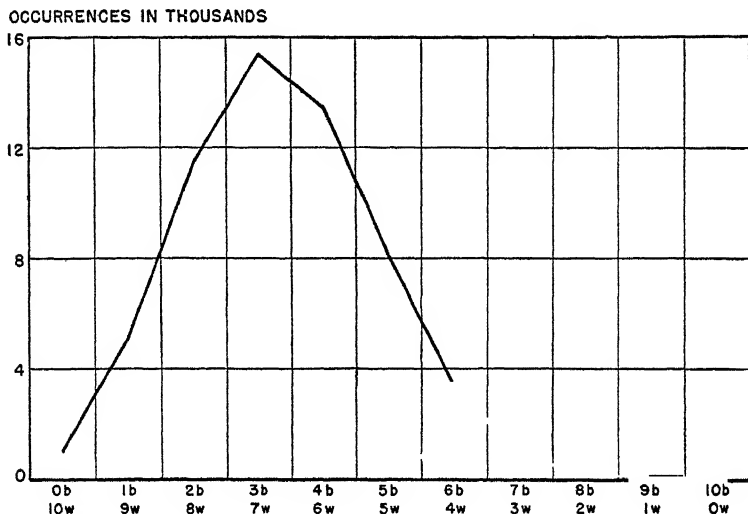
Chart 122. Expected Results of 59,049 Throws of 10 Dice, Each Having Four White

Sides and Two Black Sides. The expected occurrences are given by $\left(\dfrac{2}{3}w + \dfrac{1}{3}b\right)^{10}$

$$=\frac{1,024}{59,049}w^{10} + \frac{5,120}{59,049}bw^9 + \frac{11,520}{59,049}b^2w^8 + \frac{15,360}{59,049}b^3w^7 + \frac{13,440}{59,049}b^4w^6 + \frac{8,064}{59,049}b^5w^5$$

$$+\frac{3,360}{59,049}b^6w^4 + \frac{960}{59,049}b^7w^3 + \frac{180}{59,049}b^8w^2 + \frac{20}{59,049}b^9w + \frac{1}{59,049}b^{10}.$$

If three such dice are thrown, the expression is

$$(\tfrac{2}{3}w + \tfrac{1}{3}b)^3 = \tfrac{8}{27}w^3 + \tfrac{12}{27}w^2b + \tfrac{6}{27}wb^2 + \tfrac{1}{27}b^3.$$

It will be observed that the binomial is beginning to show its skewed nature. This will be more clearly seen if we consider throwing ten dice, each with two black sides. The expression is $(\tfrac{2}{3}w + \tfrac{1}{3}b)^{10}$, which is shown graphically in Chart 122. The curve is definitely skewed as a result of the fact that the two fractions q and p are unequal.

If q is a larger fraction and p is smaller, the skewness will be even greater. Let us consider as an illustration a four-sided pyramidal die with one

black side and three white sides. It will be necessary to consider the "down" side as the one obtained at a throw. For throwing one die, the probability is $\frac{3}{4}w + \frac{1}{4}b$.

If 10 of these four-sided dice are thrown, their behavior is indicated by $(\frac{3}{4}w + \frac{1}{4}b)^{10}$. The expansion of this binomial is shown in Chart 123, which is noticeably more skewed than the curve of Chart 122.

A Four-Sided Die, Each Side of Which Is an Equilateral Triangle.

Fitting a binomial. It is apparent from the expression for a binomial that it is a device most useful for fitting to discrete data. In order to fit a binomial to a series of observed data, the following three steps are
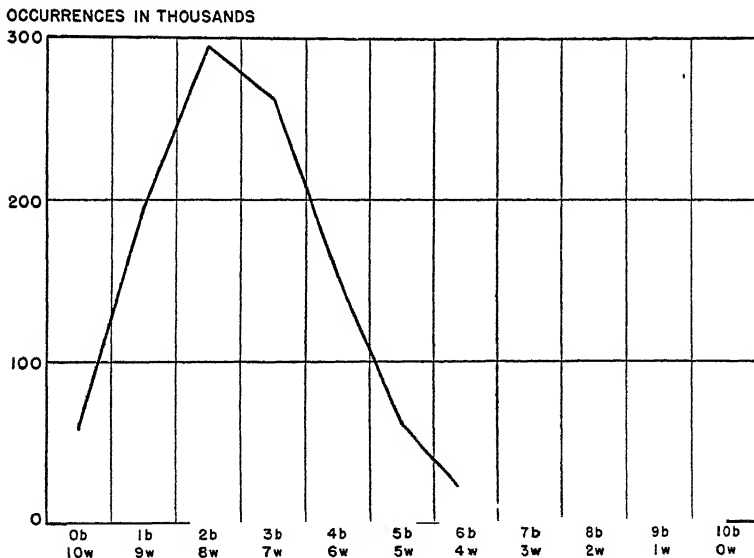
Chart 123. Expected Results of 1,048,576 Throws of 10 Four-Sided Dice, Each Having Three White Sides and One Black Side. The expected occurrences are given by

$$\left(\frac{3}{4}w + \frac{1}{4}b\right)^{10} = \frac{59,049}{1,048,576}w^{10} + \frac{196,830}{1,048,576}bw^9 + \frac{295,245}{1,048,576}b^2w^8 + \frac{262,440}{1,048,576}b^3w^7$$
$$+ \frac{153,090}{1,048,576}b^4w^6 + \frac{61,236}{1,048,576}b^5w^5 + \frac{17,010}{1,048,576}b^6w^4 + \frac{3,240}{1,048,576}b^7w^3 + \frac{405}{1,048,576}b^8w^2$$
$$+ \frac{30}{1,048,576}b^9w + \frac{1}{1,048,576}b^{10}.$$

necessary: (1) Determine the proper value of p, which also gives us q, since $q = 1 - p$. The size of p determines the degree of skewness of the curve. If $p = .50$, then $q = .50$ and the curve is symmetrical. The farther removed p is from .50, in either direction, the greater the skewness. If $p < .50$, the curve is positively skewed; if $p > .50$, it is negatively skewed. (2) Expand the binomial $(q + p)^m$, where $m =$ the number of

categories minus one, since there are $m + 1$ terms in the expanded binomial. (3) Multiply the total frequencies N by each of the fractions of the expanded binomial.

Table 62 shows a distribution of the number of male pigs occurring in litters of five pigs. The arithmetic mean of the series is computed in the usual manner and $\overline{X}$ is found to be 2.4397, the mean number of males per litter of five pigs. Since there are five pigs in a litter, the probability of any given pig (in a litter of five) being born a male is 2.4397 ÷ 5 = .4879. In similar fashion, the value of p for any such observed series is obtained from $p = \dfrac{\overline{X}}{m}$. $N = 116$, the number of litters, not 580, the number of pigs.

<div align="center">TABLE 62</div>

<div align="center">Number of Male Pigs Born in Litters of Five and
Determination of $\overline{X}$</div>

Number of males X	Number of litters having specified number of males f	fX
0	2	
1	20	20
2	41	82
3	35	105
4	14	56
5	4	20
Total	116	283

Source A S Parkes, "Studies on the Sex-Ratio and Related Phenomena The Frequencies of Sex Combinations in Pig Litters." *Biometrika*, Vol 15 (1923), pp 373–381 Parkes fits a binomial to the same series using $p = .4876$, as determined for litters of 4 to 12 pigs. His expected frequencies are identical with ours

$$\overline{X} = \frac{283}{116} = 2.4397.$$

As pointed out above, the fitting is accomplished by expanding $N(q + p)^m$. Substituting 5 for m, but retaining the other symbols

$$N(q + p)^5 = N(q^5 + 5q^4p + 10q^3p^2 + 10q^2p^3 + 5qp^4 + p^5),$$

where the exponent of p indicates the number of males born in a litter of 5.

The numerical expression to use in fitting the binomial is $(.5121 + .4879)^5$, and since $N = 116$ we should expand $116(.5121 + .4879)^5$. This becomes

$$116[(.5121)^5 + 5(.5121)^4 (.4879) + 10(.5121)^3 (.4879)^2$$
$$+ 10(.5121)^2 (.4879)^3 + 5(.5121) (.4879)^4 + (.4879)^5].$$

The computations are most conveniently carried out by means of logarithms, as shown in Table 63. Although the powers could be obtained and the multiplications could be performed for this problem by the use of a calculating machine, the use of logarithms is essential when a binomial is raised to an appreciably higher power.

Chart 124 shows the observed and the expected frequencies. The observed data have been shown by means of separated bars to suggest the discrete nature of the series. If the first two and the last two classes are combined, $\chi^2 = .72$. Since there are now four classes and since two degrees of freedom were lost (the number of litters N and p were used in
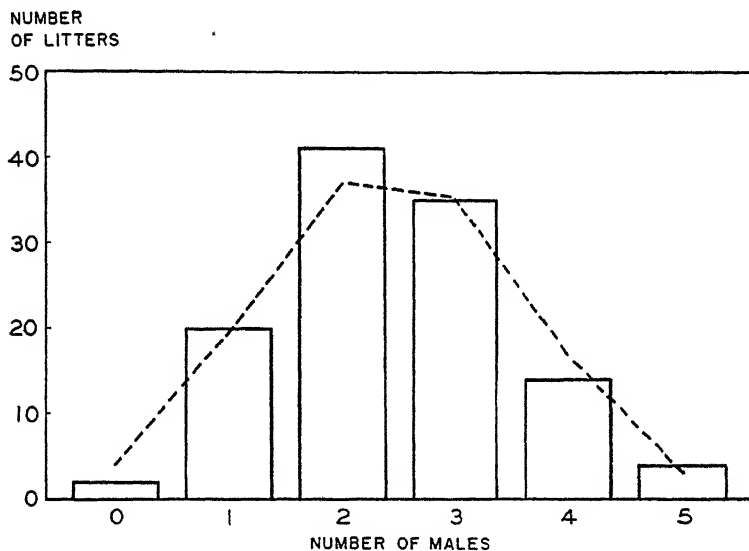
NUMBER
OF LITTERS

Chart 124. Binomial Fitted to Distribution of Number of Male Pigs Born in Litters of Five. (Data from Tables 62 and 63.)

fitting), $n = 2$. The value of P is about .70, indicating a good agreement of the observed with the expected frequencies.

It should not be assumed that all discrete series may be fitted by the method just explained. Some data require other types of series, as, for example, the Poisson series, the fitting of which is described by Tippett[15] and others.

---

[15] L. H. C. Tippett, *The Methods of Statistics*, pp. 48–54, Williams and Norgate, London, 1937 (2nd Edition). See also R. A. Fisher, *Statistical Methods for Research Workers*, pp. 56–59, Oliver and Boyd, Edinburgh, 1938 (7th edition); and H. L. Rietz (editor), *Handbook of Mathematical Statistics*, Ch. VI, Houghton Mifflin, Boston, 1924.

TABLE 63

BINOMIAL $N(q + p)^m$ FITTED TO DISTRIBUTION OF NUMBER OF MALE PIGS BORN IN LITTERS OF FIVE

$(N = 116; q = 5121; p = 4879; m = 5)$

Number of males (power of p) (1)	Expression* (2)	Log N (3)	Log C (4)	Log of indicated power of q (5)	Log of indicated power of p (6)	Σ of logs [(3) + (4) + (5) + (6)] (7)	Expected frequencies $N = 116$ [antilog of (7)] (8)
0	$N \cdot C_0 \cdot q^5 \cdot p^0 = (116)\ (1)\ (.5121)^5\ (.4879)^0$	2 064458		48 546775 — 50	9 688331 — 10	611233	4 1
1	$N \cdot C_1 \cdot q^4 \cdot p^1 = (116)\ (5)\ (.5121)^4\ (.4879)^1$	2 064458	.698970	38 837420 — 40	19.376662 — 20	1 289179	19.5
2	$N \cdot C_2 \cdot q^3 \cdot p^2 = (116)\ (10)\ (.5121)^3\ (.4879)^2$	2 064458	1.000000	29 128065 — 30	29.064993 — 30	1 569185	37 1
3	$N \cdot C_3 \cdot q^2 \cdot p^3 = (116)\ (10)\ (.5121)^2\ (.4879)^3$	2 064458	1 000000	19.418710 — 20	38 753324 — 40	1 548161	35.3
4	$N \cdot C_4 \cdot q^1 \cdot p^4 = (116)\ (5)\ (.5121)^1\ (.4879)^4$	2 064458	.698970	9.709355 — 10	48 441655 — 50	1 226107	16 8
5	$N \cdot C_5 \cdot q^0 \cdot p^5 = (116)\ (1)\ (5121)^0\ (.4879)^5$	2 064458				506113	3 2
Total							116 0

* C_0, C_1, etc., are the binomial coefficients, the multipliers for each term of the binomial expansion

$$C_0 = 1, \ C_1 = m, \ C_2 = \frac{m(m-1)}{1 \cdot 2}, \ C_3 = \frac{m(m-1)(m-2)}{1 \cdot 2 \cdot 3}, \text{ etc.}$$

292

Skewed Curves

The binomials just discussed are suitable for fitting to discrete data, but are not accurate enough to use with continuous data. A fitted binomial consists of a series of ordinates erected at specific points on the X-axis (see Chart 124). If this procedure were applied to a distribution of continuous data (or to discrete data where the X units are small in relation to the class interval), we should be erecting ordinates at the mid-value of each class, instead of determining the area under a smooth curve. Obviously, the greater the number of classes, the less would be the difference between these two procedures.

There are a great many types of skewed curves which may be fitted to frequency distributions. It is the purpose of this volume, not to enter
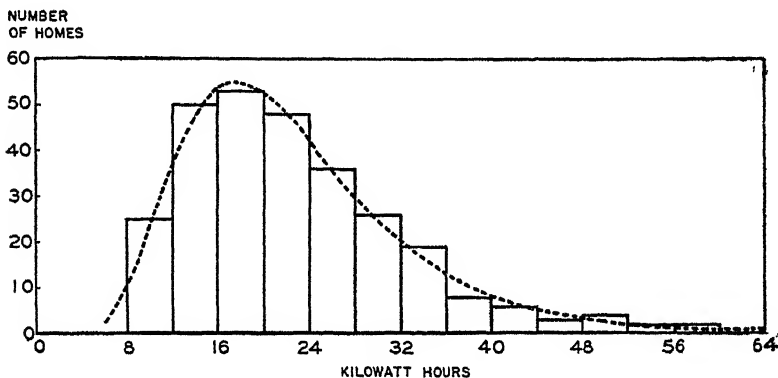
Chart 125. Logarithmic Normal Curve Fitted to Kilowatt Hours of Electricity Used per Month in 282 Medium-Class Homes in an Eastern City. (Based on data of Table 64.)

into an extended consideration of this topic, but merely to sketch briefly the procedure involved in fitting two of the simpler types.[16]

The logarithmic normal curve. Some distributions which are skewed to the right become symmetrical when plotted in terms of the logarithms of their X values or, alternately, when plotted on graph paper having a logarithmic X-scale. The column diagram of Chart 125 shows the monthly use of electricity by 282 medium class homes in an eastern city, drawn from the data of Table 64. It is apparent that the series is decidedly skewed in a positive direction. In Chart 126 these data have been re-

---

[16] For a more detailed discussion, see: W. P. Elderton, *Frequency Curves and Correlation*, Cambridge University Press, Cambridge, England, 1938 (3rd Edition); H. L. Rietz, *Mathematical Statistics*, Open Court Publishing Co., Chicago, 1927; Arne Fisher, *Mathematical Theory of Probabilities*, Macmillan, New York, 1922 (2nd Edition).

plotted but against a logarithmic X-scale. When the curve is extended to $Y = 0$ at $X = 6$ kilowatt hours (the first class below the first one shown in the table), the approximate symmetrical nature of the series

NUMBER
OF HOMES
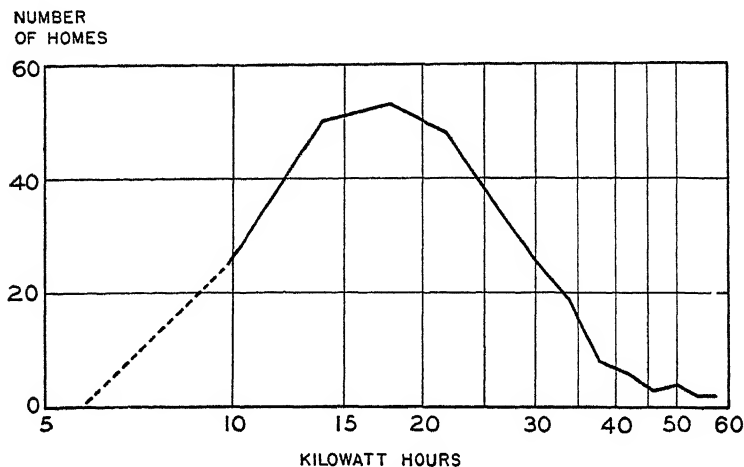
KILOWATT HOURS

Chart 126. Kilowatt Hours of Electricity Used per Month in 282 Medium-Class Homes in an Eastern City. Logarithmic X-scale. (Data of Table 64 Frequencies are plotted at logarithmic mid-values of classes.)

TABLE 64

Kilowatt Hours of Electricity Used
per Month in Medium-Class Homes
in an Eastern City

Kilowatt hours (mid-values)	Number of homes
10	25
14	50
18	53
22	48
26	36
30	26
34	19
38	8
42	6
46	3
50	4
54	2
58	2
Total	282

Source: Electrical Testing Laboratories, New York City Name of city withheld by request.

in terms of logarithmic X values is apparent. A further indication of this is shown in Chart 127, which presents the cumulative percentage frequencies plotted on logarithmic probability paper.[17]

Fitting a logarithmic normal curve. The procedure for fitting a loga

PER CENT
OF HOMES
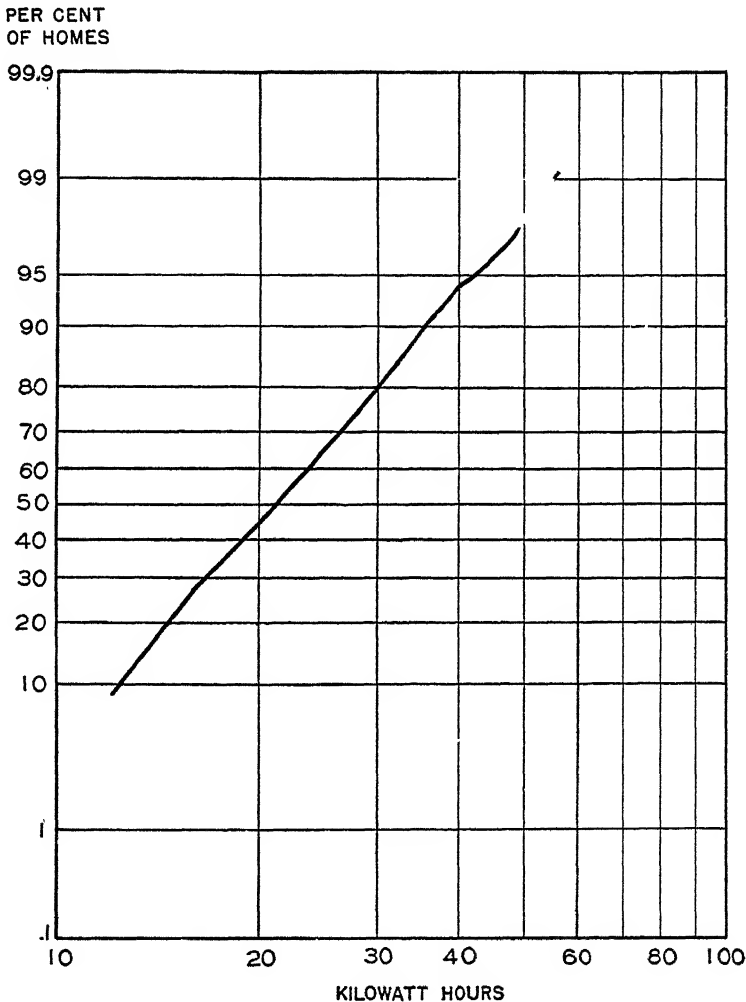
Chart 127. Kilowatt Hours of Electricity Used per Month in 282 Medium-Class Homes in an Eastern City. Shown on logarithmic probability paper. (Based on data of Table 64)

---

[17] Available from Codex Book Company, Norwood. Mass.

rithmic normal curve has been explained by Davies[18] and is essentially the same process as that of fitting a normal curve, save that we use the arithmetic mean $\overline{X}_{\log}$ and the standard deviation $\sigma_{\log}$ of the logarithms of the X values. The values of $\overline{X}_{\log}$ and $\sigma_{\log}$ may be computed by making use of the mid-values of the logarithms of the class limits. Ideally the classes should be so chosen that the class intervals are equal in a logarithmic sense, thus making the logarithmic mid-values equidistant from each other. Usually we are dealing with ready-formed frequency distributions of arithmetically equal class intervals, and with such distributions the direct computation of $\overline{X}_{\log}$ and $\sigma_{\log}$ is laborious. The inconvenience of computing these logarithmic values has been eliminated by Davies, who gives formulae based upon the quartiles which are readily computed. Furthermore, according to Davies, there are certain advantages to the procedure. He says: "Unless the data are very regular, these [$\overline{X}_{\log}$ and $\sigma_{\log}$] may be more satisfactorily computed from the quartiles, thus avoiding the disturbing effects of irregular extreme items." The expressions are given below.

$$\overline{X}_{\log} = \frac{\log Q_1 + \log Q_3 + 1.2554 \log Q_2}{3.2554}$$

This is the weighted average of the three quartiles, the weights being proportional to the heights of normal curve ordinates erected at these values.

$$\sigma_{\log} = .7413 \, (\log Q_3 - \log Q_1).$$

This expression grows out of the fact that in a normal curve 50 per cent of the items are included within $\pm Q$ of the median (or mean), and also that 50 per cent of the items are included within $\pm .6745\sigma$ of the mean. It is therefore obvious that

$$\sigma = \frac{1}{.6745} Q = 1.4825 Q.$$

Since

$$\frac{Q_3 - Q_1}{2} = Q,$$

it follows that

$$Q_3 - Q_1 = 2Q, \text{ and } \sigma = .7413(Q_3 - Q_1).$$

For the data of electric consumption, $Q_1 = 15.6400$ kwh., Q_2 (the median) $= 21.0833$ kwh., and $Q_3 = 27.9444$ kwh.

$$\overline{X}_{\log} = \frac{\log 15.6400 + \log 27.9444 + 1.2554 \log 21.0833}{3.2554}$$

---

[18] G. R. Davies and W. F. Crowder, *Methods of Statistical Analysis*, pp. 303–306; and G. R. Davies, "The Analysis of Frequency Distributions," *Journal of the American Statistical Association* Vol 24, December 1929, pp. 349–366

$$= \frac{1.194237 + 1.446295 + 1.2554(1.323939)}{3.2554}$$

$$= \frac{4.302605}{3.2554} = 1.321682.$$

$$\sigma_{\log} = .7413(\log 27.9444 - \log 15.6400)$$

$$= .7413(1.446295 - 1.194237)$$

$$= .7413\,(.252058)$$

$$= .186851$$

Using these two values, we may determine the expected frequencies in each class in a manner strictly parallel to that used previously for the normal curve and by using the same table of areas (Appendix E). Table 65 indicates the procedure. The expected frequencies and the observed frequencies are in close agreement. Note also the correspondence of the column diagram of the original data and the fitted curve in Chart 125. The ordinates are computed from the expression[19]

$$Y_C = \frac{.4343Ni}{X\,\sigma_{\log}\sqrt{2\pi}}\,e^{\frac{-x^2_{\log}}{2\sigma^2_{\log}}}.$$

Since $\sqrt{2\pi} = 2.5066$, the expression may be simplified for purposes of computation to

$$Y_C = \frac{.17326Ni}{X\sigma_{\log}}\,e^{\frac{-x^2_{\log}}{2\sigma^2_{\log}}}.$$

X is the arithmetic value of the point on the X-axis at which the ordinate

---

[19] It will be recalled that the expression for the normal curve is

$$Y_c = \frac{Ni}{\sigma\sqrt{2\pi}}\,e^{\frac{-x^2}{2\sigma^2}}$$

For fitting the logarithmic normal curve, the expression cannot be used in this form since σ is in terms of logarithms ($\sigma_{\log}$), while the class intervals i are equal arithmetically. We therefore multiply i by the adjustment factor $\frac{\log_{10} e}{X}$ or $\frac{.4343}{X}$, to compensate for the fact that the logarithms of the intervals are not equal. We thus have

$$Y_c = \frac{.4343}{X} \cdot \frac{Ni}{\sigma_{\log}\sqrt{2\pi}}\,e^{\frac{-x^2_{\log}}{2\sigma^2_{\log}}}.$$

TABLE 65

DETERMINATION OF EXPECTED FREQUENCIES FOR LOGARITHMIC NORMAL CURVE FITTED TO DATA OF KILOWATT HOURS OF ELECTRICITY USED PER MONTH IN 282 MEDIUM CLASS HOMES IN AN EASTERN CITY

$(\overline{X}_{log} = 1\ 321682;\ \sigma_{log} = 186851)$

Kilowatt hours consumed (1)	Logarithm of limits of classes — Lower limits (2)	Logarithm of limits of classes — Upper limits (3)	x_{log} (log of limit $-\ \overline{X}_{log}$) (4)	$\dfrac{x_{log}}{\sigma_{log}}$ (5)	Cumulative theoretical frequencies as percentages (Appendix E) (6)	Theoretical frequencies as percentages (7)	Theoretical frequencies $N = 282$ (8)
Below 4	..				50 00	.01	.
4 but less than 8	602060		719622	3 85	49 99	1 24	3 5
8 but less than 12	.903090		418592	2 24	48 75	8 43	23 8
12 but less than 16	1 079181		242501	1 30	40 32	16 75	47 2
16 but less than 20	1 204120		117562	63	23 57	19 19	54 1
20 but less than 24	1 301030	1 380211	020652 / 058529	11 / 31	4.38 / 12 17	16 55	46 7
24 but less than 28		1.447158	125476	67	24 86	12 69	35 8
28 but less than 32		1 505150	183468	98	33 65	8 79	24 8
32 but less than 36		1 556303	234621	1.26	39 62	5 97	16 8
36 but less than 40		1 602060	280378	1 50	43 32	3 70	10 4
40 but less than 44		1 643453	321771	1 72	45 73	2 41	6.8
44 but less than 48		1 681241	359559	1 92	47 26	1 53	4 3
48 but less than 52		1 716003	394321	2.11	48 26	1 00	2 8
52 but less than 56		1.748188	426506	2 28	48 87	.61	1 7
56 but less than 60		1 778151	456469	2.44	49 27	40	1 1
60 but less than 64		1 806180	484498	2 59	49 52	25	7
64 but less than 68		1 832509	510827	2 73	49 68	16	5
68 or more		.	.	.	50 00	32	.9
Total......						100.00	281 9

is to be erected. The values of $e^{\frac{-x_{\log}^2}{2\sigma_{\log}^2}}$ are obtained from Appendix D,

and the $\frac{x_{\log}}{\sigma_{\log}}$ values are given by

$$\frac{x_{\log}}{\sigma_{\log}} = \frac{\log X - \overline{X}_{\log}}{\sigma_{\log}}.$$

Davies suggests a logarithmic coefficient of skewness

$$\mathrm{Sk}_{\log} = \frac{\log Q_1 + \log Q_3 - 2\log Q_2}{\log Q_3 - \log Q_1}$$

and points out that a series which yields a coefficient of less than 0.15 (or perhaps even 0.20) may tentatively be considered as logarithmically normal. If, however, a skewed distribution is not inherently logarithmic, Davies notes that it may sometimes be adjusted by shifting the X values until the desired skewness is obtained; after fitting, the X values are again shifted. This correction c is obtained by

$$c = \frac{Q_2^2 - Q_1 Q_3}{Q_1 + Q_3 - 2Q_2}.$$

This value is added to the class limits and to the quartiles, after which $\overline{X}_{\log}$ and $\sigma_{\log}$ are computed. The fitting proceeds as in Table 65, but the shifted class limits are used. After the expected frequencies have been ascertained, the class limits are shifted back to their original values. It is obvious that this device extends the usefulness of the logarithmic normal curve.

Fitting a normal curve with adjustment for skewness. The formulae previously given for the normal curve enabled us to fit a symmetrical curve from a knowledge of $\overline{X}$, σ, and N. We have just considered one method of fitting a skewed curve. Another procedure that is useful for certain skewed distributions consists of using also a measure of skewness

$$\alpha_3 = \sqrt{\beta_1} = \sqrt{\frac{\mu_3^2}{\mu_2^3}} = \frac{\mu_3}{\sqrt{\mu_2^3}} \quad \left(\text{or } \frac{\pi_3}{\sqrt{\pi_2^3}} \text{ if Sheppard's correction is not applied}\right)$$

and thereby making a correction to the fit of a normal curve. This is sometimes referred to as a second approximation curve. The equation [20] is

$$Y_c = \frac{Ni}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}} - \left\{ \frac{Ni}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}} \left[\frac{\alpha_3}{2} \left(\frac{x}{\sigma} - \frac{x^3}{3\sigma^3} \right) \right] \right\}.$$

---

[20] The expression includes the first two terms of the Gram-Charlier series. For a further discussion, see W. A. Shewhart: *Economic Control of Quality of Manufactured Product*, pp 84–94, D. Van Nostrand, New York, 1931.

The expression preceding the minus sign is that for the normal curve, while the expression in braces represents a modification for skewness. In order to determine the expected frequencies, the above equation must be integrated. This is accomplished by the use of tables. To use these tables, we write

$$\int_0^x f(x)dx = F_1\left(\frac{x}{\sigma}\right) - \alpha_3 F_2\left(\frac{x}{\sigma}\right),$$

TABLE 66

COMPUTATION OF $\overline{X}$, σ, AND α_3 FOR DEPTH OF SAPWOOD

Depth in inches (mid-values)	f	d	fd'	$f(d')^2$	$f(d')^3$
1 0	2	−7	− 14	98	− 686
1 3	29	−6	−174	1,044	−6,264
1 6	62	−5	−310	1,550	−7,750
1 9	106	−4	−424	1,696	−6,784
2 2	153	−3	−459	1,377	−4,131
2 5	186	−2	−372	744	−1,488
2 8	193	−1	−193	193	− 193
3 1	188	0	0	0	0
3 4	151	1	151	151	151
3 7	123	2	246	492	984
4 0	82	3	246	738	2,214
4 3	48	4	192	768	3,072
4 6	27	5	135	675	3,375
4.9	14	6	84	504	3,024
5 2	5	7	35	245	1,715
5 5	1	8	8	64	512
Total	1,370	.	−849	10,339	−12,249

Source: Data from W A Shewhart, *Economic Control of Quality of Manufactured Product*, p 77, D Van Nostrand Co., New York, 1931 Courtesy of D Van Nostrand Co. Inc.

$$\nu_1 = \frac{\Sigma fd'}{N} = -.619708.$$

$$\nu_2 = \frac{\Sigma f(d')^2}{N} = 7\ 546715.$$

$$\nu_3 = \frac{\Sigma f(d')^3}{N} = -8.940876.$$

$$\overline{X} = 3.1 - [(.619708)(.3)] = 2.9141 \text{ inches.}$$

Since Sheppard's correction is not applied, we have

$$\pi_2 = \nu_2 - \nu_1^2 = 7.162677,$$
$$\pi_3 = \nu_3 - 3\nu_1\nu_2 + 2\nu_1^3 = 4.613422.$$
$$\sigma = i\sqrt{\pi_2} = .8029 \text{ inches.}$$
$$\alpha_3 = \sqrt{\beta_1} = \sqrt{\frac{\pi_3^2}{\pi_2^3}}, \text{ or } \frac{\pi_3}{\sqrt{\pi_2^3}} = +.2407.$$

where $F_1\left(\dfrac{x}{\sigma}\right)$ represents the areas of the normal curve (given in Appendix E) and $\alpha_3 F_2\left(\dfrac{x}{\sigma}\right)$ represents the modification for skewness. Values of $F_2\left(\dfrac{x}{\sigma}\right)$ are obtained from Appendix J and are then multiplied by α_3.

As an illustration of this method of fitting, we use the data of Table 66, which are shown graphically in Chart 128. The fitting procedure[21] for a second approximation curve is shown in Table 67. The values of N, $\overline{X}$, σ, and α_3 having been obtained (Table 66), the steps are as follows:

1. Make entries in columns (1) to (6) inclusive, as was done in fitting a normal curve.
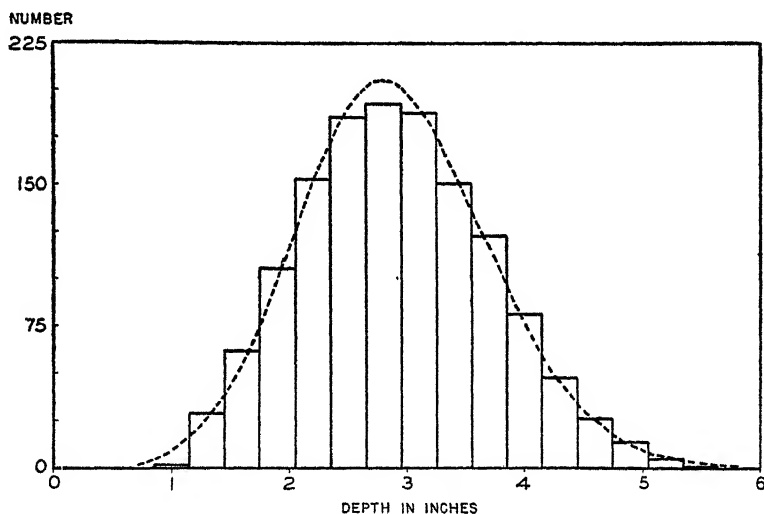
Chart 128. Second Approximation Curve Fitted to Depth of Sapwood. (Based on data of Table 66)

2. Refer to Appendix J and enter in column (7) the $F_2\left(\dfrac{x}{\sigma}\right)$ values associated with each $\dfrac{x}{\sigma}$ value of column (5). Negative signs are entered in this column for the percentages associated with class limits of column (2).

---

[21] Sheppard's correction has not been applied in the computation of the second moment, partly because the distribution is skewed. Furthermore, Shewhart points out (*op. cit.*, p. 78) that the corrected standard deviation (.798211) differs more from the standard deviation of the ungrouped data (802555) than does the uncorrected standard deviation (.802895). When high contact is not present, overcorrection of a moment is not unusual. It arises because the corrections allow for non-existent classes at the extremes.

TABLE 67

FIT OF SECOND APPROXIMATION CURVE TO DATA OF DEPTH OF SAPWOOD

($\overline{X}$ = 2.9141 inches; σ = 8029 inches; α_3 = + 2407)

Depth in inches (mid-values) (1)	Lower limits (2)	Upper limits (3)	x (4)	$\frac{x}{\sigma}$ (5)	$F_1\left(\frac{x}{\sigma}\right)$ as percentages (6)	$F_2\left(\frac{x}{\sigma}\right)$ as percentages (7)	$\alpha_3 F_2\left(\frac{x}{\sigma}\right)$ as percentages (8)	$F_1\left(\frac{x}{\sigma}\right) - \alpha_3 F_2\left(\frac{x}{\sigma}\right)$ as percentages [Col 6 − Col 8] (9)	Theoretical frequencies as percentages N = 100 per cent (10)	Theoretical frequencies N = 1,370 (11)
.4	.25		2.6641	3.318	49.95	−6.92	−1.67	51.62	.02	
.7	.55		2.3641	2.944	49.84	−7.32	−1.76	51.60	.18	2
1.0	.85		2.0641	2.571	49.49	−8.02	−1.93	51.42	.67	9
1.3	1.15		1.7641	2.197	48.60	−8.93	−2.15	50.75	1.85	25
1.6	1.45		1.4641	1.824	46.59	−9.58	−2.31	48.90	4.03	55
1.9	1.75		1.1641	1.450	42.65	−9.21	−2.22	44.87	7.23	99
2.2	2.05		.8641	1.076	35.90	−7.24	−1.74	37.64	10.77	148
2.5	2.35		.5641	.703	25.90	−4.02	−.97	26.87	13.73	188
2.8	2.65	2.95	.2641	.329	12.89	−1.03	−.25	13.14	14.93	205
			0.0359	.045	1.79	.02	—	1.79		
3.1		3.25	.3359	.418	16.21	1.62	.39	15.82	14.03	192
3.4		3.55	.6359	.792	28.58	4.84	1.16	27.42	11.60	159
3.7		3.85	.9359	1.166	37.82	7.87	1.89	35.93	8.51	117
4.0		4.15	1.2359	1.539	43.81	9.43	2.27	41.54	5.61	77
4.3		4.45	1.5359	1.913	47.21	9.49	2.28	44.93	3.39	46
4.6		4.75	1.8359	2.287	48.89	8.71	2.10	46.79	1.86	25
4.9		5.05	2.1359	2.660	49.61	7.82	1.88	47.73	.94	13
5.2		5.35	2.4359	3.034	49.88	7.20	1.73	48.15	.42	6
5.5		5.65	2.7359	3.408	49.97	6.86	1.65	48.32	.17	2
5.8		5.95	3.0359	3.781	49.99	6.72*	1.62	48.37	.05	1
6.1		6.25	3.3359	4.155*	50.00	6.67*	1.61	48.39	.02	
6.4		6.55	3.6359	4.528*	50.00	6.66*	1.60	48.40	.01	

* For values of $F_2\left(\frac{x}{\sigma}\right)$ beyond the range given in Appendix J, use the expression

$$F_2\left(\frac{x}{\sigma}\right) = \frac{1}{6\sqrt{2\pi}}\left\{1 - \left[1 - \left(\frac{x}{\sigma}\right)^2\right]e^{\frac{-x^2}{2\sigma^2}}\right\} = \frac{1}{15.036}\left\{1 - \left[1 - \left(\frac{x}{\sigma}\right)^2\right]e^{\frac{-x^2}{2\sigma^2}}\right\}$$

The values of $e^{\frac{-x^2}{2\sigma^2}}$ may be conveniently read from the table of ordinates of the normal curve (Appendix D), or from a more extensive table in Karl Pearson, *Tables for Statisticians and Biometricians*, pp 2–8, University Press, Cambridge (England), 1914. The values for z shown in the latter table yield $e^{\frac{-x^2}{2\sigma^2}}$ when multiplied by 2 5066.

3. In column (8), multiply each value of column (7) by α_3. Signs are shown.

4. To produce column (9), the values in column (8) are subtracted algebraically from the values in column (6).

5. The cumulative areas or frequencies of column (9) are decumulated in column (10), as was done for the normal curve. The result is a series of figures showing expected frequencies on the basis of the second approximation for $N = 100$ per cent. One of the shortcomings of this curve is that it may occasionally produce negative frequencies at one end, or, if we do not extend the fit far enough to produce these negative frequencies, the total may slightly exceed 100 per cent. In this instance column (10) totals 100.02.

6. In column (11) the frequencies are prorated among the classes so that the total equals the N of the sample.

Selected References

W. D. Baten: *Elementary Mathematical Statistics*, Chapter 4; John Wiley and Sons, New York, 1938. Fitting the normal curve.

F. E. Croxton and D. J. Cowden: *Practical Business Statistics*, Chapter XII, Prentice-Hall, Inc., New York, 1934 Fitting the normal curve.

G. R. Davies and W. F. Crowder: *Methods of Statistical Analysis in the Social Sciences*, pages 303–306; John Wiley and Sons, New York, 1933. The logarithmic normal curve. (See also G. R. Davies: "The Analysis of Frequency Distributions," *Journal of the American Statistical Association*, Vol. XXIV, No. 168, December 1929, pages 359–366)

W. P. Elderton· *Frequency Curves and Correlation* (Third Edition), Chapters IV, V, VI; Cambridge University Press, Cambridge, England, 1938. Curves of the Pearson system are discussed in Chapters IV and V.

R. A. Fisher: *Statistical Methods for Research Workers* (Seventh Edition), Chapters III, IV; Oliver and Boyd, Edinburgh, 1938. Normal curve, Poisson series, binomial, and χ^2 test.

F. C. Mills: *Statistical Methods Applied to Economics and Business* (Revised Edition), pages 425–448, 618–636; Henry Holt and Co., New York, 1938. The normal curve and the χ^2 test.

P. R. Rider: *An Introduction to Modern Statistical Methods*, pages 67–75, 108–110; John Wiley and Sons, New York, 1939. Normal curve and χ^2 test.

H. L. Rietz: *Mathematical Statistics*, Chapter III; Open Court Publishing Co., Chicago, 1927. Discusses Gram-Charlier curves and curves of the Pearson system.

W. A. Shewhart: *Economic Control of Quality of Manufactured Product*, pages 85–94; D. Van Nostrand Co., New York, 1931. The normal curve and modification for skewness.

G. W. Snedecor: *Statistical Methods Applied to Experiments in Agriculture and Biology*, Chapters 1, 9, 16; Collegiate Press, Ames, Iowa, 1937. An experimental approach to χ^2 is given in Chapter 1. Binomial and Poisson distributions are discussed in Chapter 16

L. H. C. Tippett: *The Methods of Statistics* (Second Edition), pages 43–62, 98–102, Williams and Norgate, London, 1937. Binomial, Poisson, and normal distributions and the χ^2 test.

A. E. Waugh· *Elements of Statistical Method*, pages 85–113; McGraw-Hill Book Co., New York, 1938. The normal curve and the Poisson series. In his *Laboratory Manual and Problems for Elements of Statistical Method*, pages 12–19 (McGraw-Hill Book Co., New York, 1938), Waugh gives tables for fitting skewed curves of the Pearson Type III. These tables were condensed from more detailed tables given by R. S. Salvosa in *Annals of Mathematical Statistics*, Vol. 1, No. 2, May 1930.

G. U. Yule and M. G. Kendall: *An Introduction to the Theory of Statistics* (Eleventh Edition), Chapters 10, 22; Charles Griffin and Co , Ltd., London, 1937. Binomial, normal, and Poisson distributions and the χ^2 test.

CHAPTER XII
RELIABILITY AND SIGNIFICANCE OF STATISTICAL MEASURES
ARITHMETIC MEANS

In the previous chapter it was observed that errors of measurement tend to follow the normal law   The chance occurrences growing out of tossing coins and from partitioning sand by means of the apparatus of Chart 114 likewise show a definite tendency to assume the normal shape.   In this and the following chapter we are interested in a study of means and other computed values, obtained from samples of a larger body of data   We shall want to know how much reliance we can place in a statistical measure computed from a sample, and we shall make use of the concept of the normal curve to begin our attack upon this problem.

Reliability of Sample Means, Large Samples

If we have a large body of data—for example, figures for thousands of automobile tires (of the same size, quality, and make, and used on similar vehicles) showing the distance run by each tire—we may study the entire set of data or we may study one or more samples drawn at random from the data.   If we select a sample of (say) 200 items, we shall find that the sample will furnish us much useful information.   For instance, the mean from the sample will probably not be greatly different from the mean of all of the data.   Furthermore, the larger the size of the sample (and the smaller the variability in the basic data), the greater the likelihood that our sample mean will agree closely with the population mean. If we selected an additional sample of 200 items, we should not necessarily get the same value for the mean of this sample as for the first.   It would be possible to select 1,000 samples of 200 items each and then, from the 1,000 means thus obtained, we could determine a standard deviation of the sample means, which would tell us the amount of variability present in our sample means.   Thus we might write

$$\sigma_{\overline{x}} = \sqrt{\frac{(\overline{X}_1 - \overline{X}_P)^2 + (\overline{X}_2 - \overline{X}_P)^2 + \cdots + (\overline{X}_k - \overline{X}_P)^2}{k}},$$

where $\sigma_{\bar{X}}$ is the standard deviation of the sample means;

$\bar{X}_1$, $\bar{X}_2$, etc., are the means of successive samples;

$\bar{X}_P$ is the mean of all the items in the population (that is, the population mean);

k is the number of means considered (or the number of samples drawn).

It would be very unusual to have such an exhaustive collection of sample data as indicated above. Generally there is available but one or a few samples of a larger group, commonly referred to as the "population" or "universe." It so happens, however, that the means of samples drawn

PER CENT
OF TOTAL
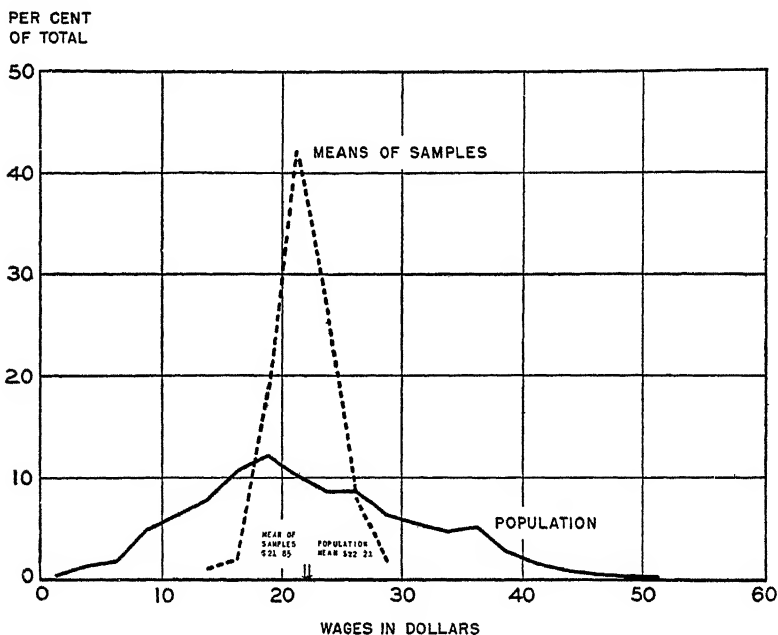
Chart 129. Distribution of Means of 100 Random Samples of 10 Items Each and of Population of 972 Wage Earners' Weekly Earnings. Even though each sample consisted of but 10 items, the approximate symmetry of the curve of sample means is apparent.

at random from a normal population tend to form a normal curve around the population mean, and for that reason it is easily possible to infer, from limited information, the behavior of arithmetic means computed from such samples. If the distribution of the population is not exactly normal, the distribution of means of random samples tends to normality as the size of the sample is increased. (The behavior of measures computed from small samples will be discussed in the latter part of this chapter.)

Chart 129 shows the results of drawing a number of samples, each of 10 items, from a larger population. The solid curve shows the original distribution which is definitely skewed. The curve of the means, however, is nearly symmetrical. Notice the more limited spread of the curve of the sample means. If the size of the 100 samples had been larger, the spread of the curve of sample means would have been less.[1]

The standard error of a sample mean. Suppose that, for the thousands of tires referred to previously, the mean mileage is 15,200 and the standard deviation is 1,248 miles. Suppose, further, that a sample of 900 tires shows $\bar{X}$ = 15,223 miles and σ = 1,230 miles. We are interested in knowing, first, how much variability should be expected in means from samples of this size and, second, whether or not 15,223 miles represents a significant divergence from the population mean. The variability may be obtained from[2]

$$\sigma_{\bar{x}} = \frac{\sigma_P}{\sqrt{N}},$$

where $\sigma_{\bar{x}}$ is the standard error of means drawn from samples (the value which would be obtained if we should compute the standard deviation of the means of all possible samples of N items);

σ_P is the standard deviation of the entire population;

N is the number of items in the sample.

For the above data

$$\sigma_{\bar{x}} = \frac{1248}{\sqrt{900}} = 41.6 \text{ miles.}$$

---

[1] To select a random sample experimentally, we may record all of the original (population) data on small cardboard discs and place them in a container, then mix the discs thoroughly, select one disc, record the entry, replace the disc, mix again, and repeat.

[2] For a derivation of this expression, see Appendix B, section XII–1 As is apparent from that development, our methods assume that the population is infinite. In Fisher's words, "the values or sets of values before us are interpreted as a random sample of a hypothetical infinite population of such values as might have arisen in the same circumstances." (R. A. Fisher, *Statistical Methods for Research Workers*, p 7, Oliver and Boyd, Edinburgh, 1938, 7th Edition). If the population is finite, but very large, the expression $\sigma_{\bar{x}} = \dfrac{\sigma_P}{\sqrt{N}}$ is adequate. However, if the population is finite and the number in the sample N is not negligible in relation to the size of the population P then, as is shown in Appendix B, section XII–1, the expression becomes

$$\sigma_{\bar{x}} = \frac{\sigma_P}{\sqrt{N}} \sqrt{\frac{P - N}{P - 1}}.$$

Sample medians are less reliable than sample means; $\sigma_{\text{med}} = 1.2533\sigma_{\bar{x}}$ if the population is normal. For a discussion of the reliability of the median, quartiles, and percentiles, see G. Udny Yule and M. G. Kendall, *An Introduction to the Theory of Statistics*, pp. 380–385, Charles Griffin and Co., London, 1937 (11th Edition).

The interpretation of this measure is analogous to that of the standard deviation. If additional samples of the same size are drawn at random from this population, we should expect 68.27 per cent of the means to fall within ±41.6 miles of the population mean, that is, within the range of 15,241.6 and 15,158.4 miles; we should expect 95.45 per cent to fall within a range of 15,200 ± 83.2 miles (population mean ±$2\sigma_{\bar{x}}$); we should expect 99.73 per cent or nearly all to fall within 15,200 ± 124.8 miles (population mean ±$3\sigma_{\bar{x}}$). For convenience we occasionally use a measure known as the probable error of the mean (PE$_{\bar{x}}$); this is[3] .6745$\sigma_{\bar{x}}$. Plus and minus one PE$_{\bar{x}}$ of the population mean indicates the range within which 50 per cent of the sample means would be expected to fall. Thus 50 per cent of the sample means would be expected to fall within 15,200 ± 28.1 miles. It will be noticed that the sample mean shown above is well within the 50 per cent range.

Significance of the deviation of a sample mean from the mean of a known population. The figures just given tell us what variation may be expected in sample means, because of the operation of chance in the drawing of random samples. We know that 68.27 per cent of the sample means would be expected to fall between 15,241.6 and 15,158 4 miles. The mean which we obtained from the sample of 900 cases was 15,223 miles. This differs +23 miles from the population mean. What is the probability of obtaining a sample mean differing by +23 miles or more from the population mean? Chart 130 shows graphically that the probability of getting a sample mean which differs by +23 miles or more from the population mean is very large. We may put this in numerical terms by determining the area of the cross-hatched section shown on the curve. We take the distance on the horizontal scale from the population mean to the observed mean as x; thus

$$x = \bar{X} - \bar{X}_P = 15{,}223 - 15{,}200 = +23 \text{ miles.}$$

If we express this deviation in terms of $\sigma_{\bar{x}}$, we may ascertain the area of the white section A of Chart 130 from Appendix E. Thus

$$\frac{x}{\sigma} = \frac{\bar{X} - \bar{X}_P}{\sigma_{\bar{x}}} = \frac{23}{41.6} = .55,$$

and from the appendix we find that 20.88 per cent of the total area is included between 15,223 and 15,200 on the horizontal scale. Subtracting this from 50 per cent (half the curve lies above the mean), we have 29.12 per cent, indicating that in 29 cases out of 100 we might expect a sample

---

[3] See Appendix E. Twenty-five per cent of the area of a normal curve is included between an ordinate erected at the mean and an ordinate erected at a distance .6745σ on the horizontal axis from the mean.

mean to *exceed* the population mean by 23 miles or more.   Sometimes we wish to ascertain the probability that a sample mean might *either* exceed or fall below the population mean by a given amount.   Suppose we wish to ascertain the chances that a sample mean might *either* exceed or fall below the population mean by 23 miles.   This is $29 + 29 = 58$ chances out of 100 (graphically, the addition of the cross-hatched and stippled parts of Chart 130), and it is apparent that the chance variations of sampling may have caused the variation of 23 miles.   This difference there-
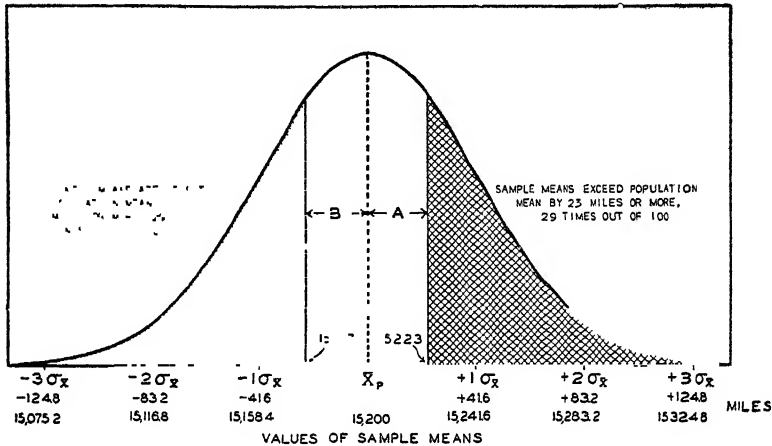
Chart 130.   Expected Distribution of Sample Means of Tire Mileage and Chances of Obtaining Sample Means Differing from the Population Mean by $+23$ and -23 Miles, When $\bar{X}_P = 15,200$ Miles, $N = 900$, and $\sigma_{\bar{x}} = 41.6$ Miles.

fore is not significant, indicating that the sample may well have been a random sample from the known population

Suppose that a sample mean ($N = 900$), possibly taken from the same universe, was 15,071 miles.   This differs from the population mean by -129 miles.   What are the chances that a mean of a random sample might differ by -129 miles or more from the population mean?   The value of $\sigma_{\bar{x}}$ is 41.6 miles, as before, and

$$\frac{x}{\sigma} = \frac{\bar{X} - \bar{X}_P}{\sigma_{\bar{x}}} = \frac{129}{41.6} = 3.10$$

From Appendix E we find that $\dfrac{4,990.3}{10,000}$ of the curve is included between 15,200 and 15,071 miles, therefore $\dfrac{9.7}{10,000}$ is included to the left of 15,071. This is Area I of Chart 131.   By pure chance, sample means would fall below the population mean 5,000 times out of 10,000.   Since there are

but 9.7 chances out of 10,000, or about 1 out of 1,000, that a sample mean might fall below the population mean by 129 miles or more, we must conclude that the difference is real.

Adding to Area I of Chart 131 the part designated as Area II, which shows the chances that a sample mean might *exceed* the population mean by 129 miles or more, we have $\frac{19.4}{10,000}$, or 2 per 1,000, as the chances that a sample mean ($N = 900$) might *exceed or fall below* the population mean by 129 miles or more. Upon either this or the preceding basis, chance is
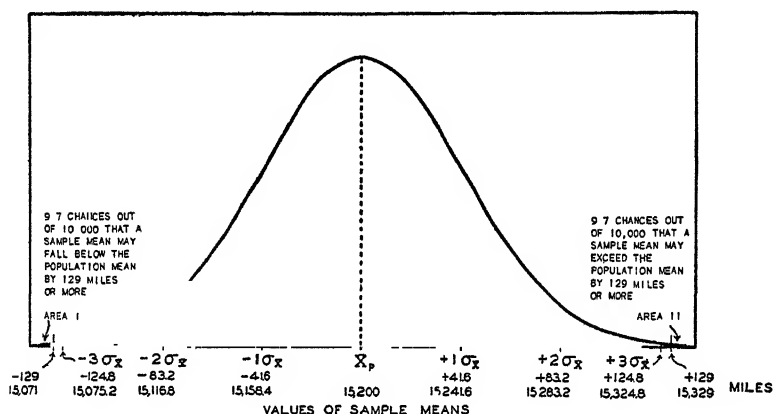
Chart 131. Expected Distribution of Sample Means of Tire Mileage and Chances of Obtaining Sample Means Differing from the Population Mean by $+129$ Miles and -129 Miles, When $\overline{X}_P = 15,200$ Miles, $N = 900$, and $\sigma_{\overline{x}} = 41.6$ Miles.

virtually ruled out, the sample mean is significantly different from the population mean, and we must conclude either that the sample was actually drawn from a different universe or that the sample was improperly chosen.

The null hypothesis. We have just set up the hypothesis that our sample of 900, which has a mean of 15,071 miles, is a random sample drawn from the population having a known mean of 15,200 miles. We then proceeded to ascertain the probability that a difference as great as that between the population mean and the observed sample mean might occur, because of chance factors arising from random sampling. The difference was so great that much doubt was cast upon our hypothesis and we abandoned it, concluding that the sample mean was significantly different from the population mean.

Such a hypothesis is called a *null* hypothesis, since our experiment or computations undertake to nullify it. The hypothesis is never proved;

neither is it disproved.   Our inference merely casts much doubt upon it (thereby impugning it) or casts little doubt upon it.

In our further study of significance of differences we shall consider the significance of the difference between a sample value and an assumed population value, and the significance of the difference between two sample values.   The procedure for testing the significance of the difference may be summarized into three steps: (1) Set up the hypothesis that the true difference is zero (i.e., that the sample has been drawn from the known or assumed population or that the two samples were drawn from the same population).   (2)   Upon the basis of this hypothesis, determine the probability that such a difference as the one observed might occur because of sampling variations.   (3)   Draw a conclusion concerning the reasonableness of the hypothesis.   If such an observed difference could hardly have occurred by chance, we have cast much doubt upon the hypothesis of (1). We therefore abandon the hypothesis and conclude that the observed difference is significant.   However, if such an observed difference could very often occur because of chance, we have cast very little doubt upon the hypothesis.   We therefore continue to regard the hypothesis as tenable and conclude that the difference is not significant.

Reliability of a sample mean, when $\overline{X}_P$ and σ_P are unknown.   The illustration just discussed assumed that the mean and the standard deviation of the population were known.   Ordinarily these population values are not known.   Our actual knowledge is often limited to the values computed from one (or a few) samples.[4]

It will be recalled that the standard error of a mean is computed by referring to the standard deviation of the universe and the number of items in the sample.   Since we do not know the standard deviation for the population, we estimate it from the sample.   While sample means vary around the population mean and may be either larger or smaller than the population mean, a sample standard deviation tends to be smaller than the standard deviation of the population.[5]   The standard deviation of the population is estimated from the sample by the expression[6]

$$\bar{\sigma} = \sqrt{\frac{\Sigma x^2}{N-1}},$$

---

[4] Occasionally a population mean (or other measure) may be obtained by setting up a control population.   In manufacturing, for example, this is accomplished by producing a large number of units under carefully controlled conditions.

[5] The standard deviation of the sample is computed in relation to its own mean. The sum of the squares of a set of deviations Σx^2 is a minimum when taken around their mean.   If these deviations were computed in relation to the population mean (be it larger or smaller than the sample mean), the sum of their squares would be greater. (See L. D. H. Weld, *Theory of Errors and Least Squares*, p. 161, Macmillan, New York, 1916.)

[6] See Appendix B, section XII-2, for a development of this expression.

where $\bar{\sigma}$ is the standard deviation of the population as *estimated from the sample*;

Σx^2 is the sum of the squared deviations of each item in the sample from the sample mean;

N is the number of items in the sample.

The expression $N - 1$ is a particular statement of the degrees of freedom (n) present. When the arithmetic mean is computed, one degree of freedom is lost, since the value of any one of the items is defined by knowledge of the value of the mean and of the remaining items. In other words, if for a series of items the sole requirement set up is that $\Sigma x = 0$ (that is, the mean has been determined), the values of all of the other items save one may be arbitrarily set down; all but one are "free to vary." The value of the other item is determined by the above requirement. In more general language, "the number of degrees of freedom is the number of deviations [items or N] minus the number of constants determined from the sample and used to fix the points from which those deviations are measured."[7] The use of $N - 1$ in the above expression is particularly important when N is small. When N is large, it matters little whether we divide by N or $N - 1$.

For purposes of computation the expression may be put in the following forms:[8]

For ungrouped data:

$$\bar{\sigma} = \sqrt{\frac{\Sigma X^2}{N - 1} - \frac{(\Sigma X)^2}{N(N - 1)}}, \text{ or } \sqrt{\frac{N\Sigma X^2 - (\Sigma X)^2}{N(N - 1)}}, \text{ or } \sqrt{\frac{\Sigma X^2 - \bar{X}\Sigma X}{N - 1}}.$$

For grouped data:

$$\bar{\sigma} = i\sqrt{\frac{\Sigma f(d')^2}{N - 1} - \frac{(\Sigma f d')^2}{N(N - 1)}}, \text{ or } i\sqrt{\frac{N\Sigma f(d')^2 - (\Sigma f d')^2}{N(N - 1)}}, \text{ or }$$

$$i\sqrt{\frac{\Sigma f(d')^2 - \dfrac{(\Sigma f d')^2}{N}}{N - 1}}.$$

---

[7] See L. H. C. Tippett, *The Methods of Statistics*, pp. 110–111, Williams and Norgate, London, 1937 (2nd Edition).

[8] For derivation, see Appendix B, section XII-3. When σ has been computed for a series of data and it is desired to transform it to $\bar{\sigma}$, use the expression

$$\bar{\sigma}^2 = \frac{N}{N - 1}\sigma^2;$$

since multiplying σ^2 by N gives Σx^2. Note that this transformation may be made even when we do not have either the observed values of each item or a frequency distribution.

In Table 68, data are shown of the weights at time of demobilization in 1919 of 746 soldiers of French extraction in the United States Army.   The computations of $\overline{X}$, σ, and $\bar{\sigma}$ appear below the table, and it will be noticed that there is but little difference between σ and $\bar{\sigma}$ because N is large.   The

TABLE 68

WEIGHT AT DEMOBILIZATION OF 746 FRENCH* SOLDIERS SERVING IN THE UNITED STATES ARMY DURING THE WORLD WAR

Weight in pounds	f	d'	fd'	$f(d')^2$
100 and under 110	7	-4	$-$ 28	112
110 and under 120	39	-3	-117	351
120 and under 130	123	-2	-246	492
130 and under 140	181	-1	-181	181
140 and under 150	183	0	0	0
150 and under 160	122	$+1$	122	122
160 and under 170	59	$+2$	118	236
170 and under 180	19	$+3$	57	171
180 and under 190	5	$+4$	20	80
190 and under 200	5	$+5$	25	125
200 and over†	3	$+6$	18	108
Total . .	746		-212	1,978

* Soldiers classified as French were either (1) born in France, or (2) had parents who were both born in France, or (3) had three or four grandparents born in France
† Mid-value taken as 205 pounds, giving results in agreement with those of the source
Source. Charles B Davenport and Albert G Love, *The Medical Department of the United States Army in the World War*, Vol XV, Part 1, pp 60 and 135 United States Government Printing Office, Washington, 1921

$$\overline{X} = 145 - \left(\frac{212}{746}\right) 10 = 142.16 \text{ pounds.}$$

$$\sigma = 10 \sqrt{\frac{1978}{746} - \left(\frac{212}{746}\right)^2} = 16.03 \text{ pounds.}$$

$$\bar{\sigma} = i \sqrt{\frac{\Sigma f(d')^2}{N - 1} - \frac{(\Sigma fd')^2}{N(N - 1)}} = 10 \sqrt{\frac{1978}{745} - \frac{(212)^2}{746 \cdot 745}} = 16.04 \text{ pounds.}$$

standard error of the mean is computed by using $\bar{\sigma}$ in place of σ_P, since the latter is unknown:[9]

$$\sigma_{\overline{X}} = \frac{\bar{\sigma}}{\sqrt{N}}.$$

---

[9] We may also obtain the standard error of the mean from

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{N - 1}},$$

since $\dfrac{\bar{\sigma}}{\sqrt{N}} = \dfrac{\sqrt{\dfrac{\Sigma x^2}{N - 1}}}{\sqrt{N}} = \sqrt{\dfrac{\Sigma x^2}{N(N - 1)}} = \dfrac{\sqrt{\dfrac{\Sigma x^2}{N}}}{\sqrt{N - 1}} = \dfrac{\sigma}{\sqrt{N - 1}}.$

For the soldiers of French extraction,

$$\sigma_{\bar{X}} = \frac{16.04}{\sqrt{746}} = \frac{16.04}{27.31} = .587,$$

which states that 2 times out of 3, sample means would fall within a range of $\pm.587$ pounds ($\pm\sigma_{\bar{X}}$) of the population mean; that 95 times out of 100, sample means would fall within ±1.174 pounds ($\pm2\sigma_{\bar{X}}$) of the population mean; and that 99.7 times out of 100, sample means would fall within ±1.761 pounds ($\pm3\sigma_{\bar{X}}$) of the population mean.

Assuming that the sample mean is smaller than the population mean by $3\sigma_{\bar{X}}$, the population mean would be 143.92 pounds. Assuming that the sample mean is larger than the population mean by $3\sigma_{\bar{X}}$, the population mean would be 140.40 pounds. Since 99.7 per cent of the sample means vary from the population mean by not more than $\pm3\sigma_{\bar{X}}$, we may conclude that the mean weight of all such soldiers in the United States Army is almost surely not less than 140.40 pounds or more than 143 92 pounds. It should be noted that, while we are able to state the probability that sample means may fall within a given range around the population mean, it is not possible to give a statement of the *probability* that the population mean falls within a given range of the sample mean, since there can be no such thing as a distribution of population means around the sample mean. We can say, however, that, if many such statements as the one concerning $\bar{X}_P$ are made in regard to this (or some other) population, we should expect to be correct in about 99.7 per cent of the instances. Statisticians make use of the concept of *fiducial probability* to express their confidence that the population mean falls within given limits. Fiducial (or fiduciary) probability states a degree of reasonable expectation and should not be confused with the mathematical probabilities referring to the variations present in statistical measures computed from samples. For the soldiers of French extraction, therefore, we may say that the fiducial probability is a little over 95 per cent that the population mean lies between 140.99 and 143.33 pounds; this is the mean of the sample (142.16 pounds) plus and minus $2\sigma_{\bar{X}}$. Similarly, the fiducial probability is 997 out of 1,000 that the population mean lies between 140.40 and 143.92 pounds ($\bar{X} \pm 3\sigma_{\bar{X}}$). Fiducial probabilities, which we have just used to state the "fiducial limits" or "confidence limits" within which a population mean might fall, must not be regarded as exact statements concerning the probability that $\bar{X}_P$ falls within given limits but rather as an expression of the degree of confidence which the statistician has in his conclusions.

Significance of the difference between a sample mean and a hypothetical population mean. Tests were made of a sample of 50 pieces of a certain

type of steel wire and revealed a mean strength of 1,221 pounds and $\bar{\sigma}$ of 49 28 pounds. These wires are to be used for "spinning" wire cable and it is essential that the mean value of the population from which the sample was taken should be *at least* 1,215 pounds. We wish to assure ourselves, then, that the sample mean is significantly greater than 1,215 pounds. What are the chances that such an observed mean may exceed the population mean by 6 pounds or more? Note that in this instance we are interested in knowing the chances that such an observed mean may *exceed* the population mean by 6 pounds, not the chances that such an observed mean may differ from (either exceed or fall below) the population mean by 6 pounds. Obviously the latter would have twice the probability of the former. The value of $\sigma_{\bar{X}}$ is

$$\frac{\bar{\sigma}}{\sqrt{N}} = \frac{49.28}{\sqrt{50}} = \frac{49.28}{7.07} = 6.97 \text{ pounds.}$$

This figure indicates the dispersion of sample means about the population mean, *not* the spread of sample means around the given sample mean. The population mean is unknown, and so we shall proceed by assuming it to be 1,215 pounds and shall ascertain if a variation as great as 6 pounds could occur in a sample by chance. This difference is .86 times the value of $\sigma_{\bar{X}}$. Since the sampling distribution of means is essentially normal, we look up .86 in Appendix E, which gives the areas of the normal curve. The tabled value .3051 indicates that there are 19 5 chances (.5000 − .3051 = .1949) out of 100 that a sample mean may have a value of 1,221 pounds or more, if the population mean is 1,215 pounds. The chances of such an occurrence being this large, we conclude that the population mean may quite possibly be as low as 1,215 pounds. The process of reasoning which we have just discussed is shown graphically in section A of Chart 132.

The above does not give us reason to conclude that the difference of 6 pounds between the sample mean and the assumed population mean is significant. The difference might become significant if we enlarge the sample. If, now, $N = 400$ while $\bar{X}$ remains 1,221 pounds and $\bar{\sigma}$ is 49.28 pounds, as before, we have

$$\frac{\bar{\sigma}}{\sqrt{N}} = \frac{49.28}{\sqrt{400}} = 2.46 \text{ pounds.}$$

The ratio of the difference (6 pounds) to the standard error of the mean is $\frac{6}{2.46} = 2.44$, which indicates that there are 73 chances out of 10,000 (or just under 1 in 100) that a sample mean may have a value of 1,221 pounds or more, if the population mean is 1,215 pounds. This is shown in section

B of Chart 132. The chances of such an occurrence being this small, we conclude that there is a significant difference and that the population mean is unlikely to be so low as 1,215 pounds. The above two problems could
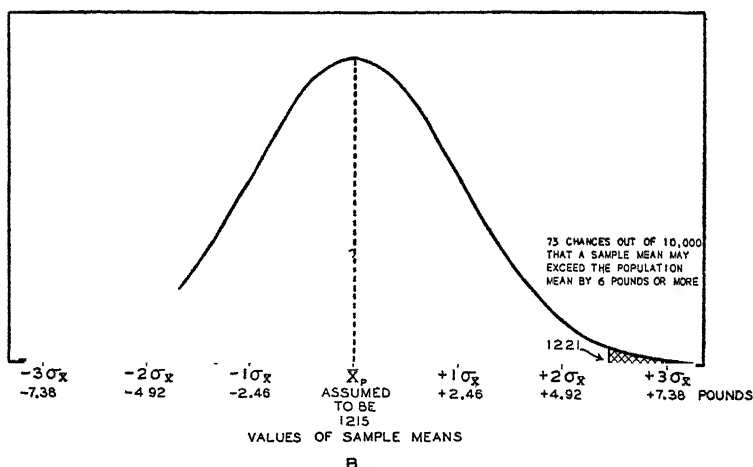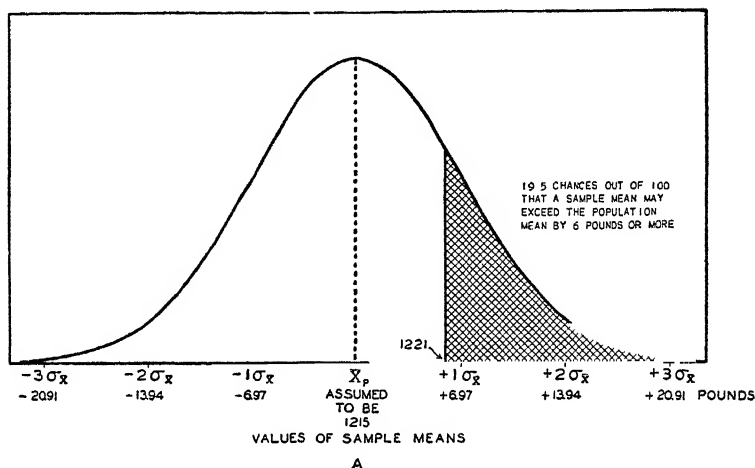
Chart 132. Expected Distribution of Sample Means of Tensile Strength of Steel Wire and Chances of Obtaining Sample Means Differing from the Population Mean by +6 Pounds: A, When $N = 50$, $\overline{X} = 1221$ Pounds, $\hat{\sigma} = 49.28$ Pounds, $\sigma_{\overline{X}} = 6.97$ Pounds; B, When $N = 400$, $\overline{X} = 1221$, $\hat{\sigma} = 49.28$, $\sigma_{\overline{X}} = 2.46$ Pounds.

have been attacked by use of fiducial probability and the conclusions would have been the same.

In the illustration just discussed, we found that 73 times out of 10,000 an event might occur owing to chance and we concluded that chance was

therefore ruled out as an explanation. Just how small should the probabilities be before we regard chance as having been virtually eliminated as a cause?

Some authorities recommend that the difference between the population mean and the observed mean should be two times the standard error of the sample mean. If we are testing whether a value may *either exceed or fall below* the population mean, the criterion of "twice the standard error" indicates that such a difference may occur, through chance, 454 times out of 10,000 (see Appendix E) This is 4.54 chances out of 100 and is often referred to as the ".05 level of significance." More accurately, the .05 level is at $\pm 1.96\,\sigma_{\bar{x}}$. It must be obvious that, if we are computing the probability that this observed sample mean might *exceed* the population mean by an amount equal to 1.96 times the standard error or more, the chances are 250 out of 10,000 (2.50 out of 100). Similarly, the chances that the sample mean might *fall below* the population mean by as much as $1.96\sigma_{\bar{x}}$ or more are 250 out of 10,000. Since it cannot be held that a sample mean is significantly *greater* than the population mean unless the sample mean is also significantly *different* from the population mean, it would seem to follow that the $\dfrac{x}{\sigma}$ ratio which is required to establish a significant difference in one direction only is the same ratio that is required to show a difference in either direction.

For a more rigid interpretation, others suggest that the observed difference should be 2.58 times its standard error. If this is so, then (from Appendix E) the probability is 98 out of 10,000, or about 1 out of 100, that the [difference (either $+$ or $-$) might have occurred by chance. This is referred to as the ".01 level of significance." An observed sample mean might *exceed* the population mean by $2.58\sigma_{\bar{x}}$, or more, .49 times out of 100; and might *fall below* the population mean $2.58\sigma_{\bar{x}}$, or more, .49 times out of 100.

Perhaps most satisfactory of all is to ascertain the probability that an observed sample mean might occur because of chance, and then to *decide whether or not the probability is small enough for the particular problem at hand.* If a test is run of the strength of window-sash cord (used to connect the upper or lower sash of a window and the sash weight), the investigator would doubtless be satisfied if the probability of obtaining a sample value as far above a specified minimum standard as that encountered were 5 in 100. On the other hand, if tests are made of the strength of parachute cord, the probability of the observed divergence should be much less. The failure of window-sash cord involves inconvenience and expense; the failure of parachute cord means tragedy.

Significance of the difference between two sample means. The intelligence quotient (I.Q.) ratings of a group of 68 left-handed students showed

a mean of 110.52. A similar group of 68 right-handed students[10] had a mean I.Q. of 109.48. Are the left-handed students actually above the right-handed in respect to I.Q. rating, or is the difference so slight as perhaps to have been due to chance variations arising in sampling? For left-handed students, $N_1 = 68$, $\overline{X}_1 = 110.52$, $\bar{\sigma}_1 = 15.2$. For right-handed students, $N_2 = 68$, $\overline{X}_2 = 109.48$, $\bar{\sigma}_2 = 15.5$.

One way to attack this problem would be to compute $\sigma_{\overline{X}_1}$ for the left-handed group and to ascertain, by fiducial probabilities, how much lower than 110.52 the population mean might be (for example, $110.52 - 3\sigma_{\overline{X}_1}$) Then compute $\sigma_{\overline{X}_2}$ for the right-handed group and determine, by fiducial probabilities, how much above 109.48 the population mean for these students might be (for example, $109.48 + 3\sigma_{\overline{X}_2}$). If the first value still exceeds the second, we may be reasonably certain that a real difference exists. But, suppose the two values just meet, or overlap slightly. Difficulties of interpretation immediately arise.

A much better procedure consists of determining the value of the standard error of the difference between the two sample means and comparing this with the observed difference between the sample means. The standard error of the difference between the two means $\sigma_{\overline{X}_1 - \overline{X}_2}$ is given by [11]

$$\sigma_{\overline{X}_1 - \overline{X}_2} = \sqrt{\sigma_{\overline{X}_1}^2 + \sigma_{\overline{X}_2}^2}.$$

For the left-handed students:

$$\sigma_{\overline{X}_1} = \frac{15.2}{\sqrt{68}} = \frac{15.2}{8.246} = 1.84.$$

For the right-handed students:

$$\sigma_{\overline{X}_2} = \frac{15.5}{\sqrt{68}} = \frac{15.5}{8.246} = 1.88.$$

[10] Based on data from Ralph Haefner, *The Educational Significance of Left-Handedness*, p. 28, Teachers College, Columbia University, New York, 1929.

[11] See Appendix B, section XII-4, where it is shown that $\sigma_{\overline{X}_1 - \overline{X}_2} = \sqrt{\sigma_{\overline{X}_1}^2 + \sigma_{\overline{X}_2}^2}$ provided there is no correlation between paired sample means. This will tend to be the situation when there is no inherent pairing between the *items* of the two series being compared.

If pairing exists between the items of the two samples being compared and there is correlation between these paired items, it may or may not be true that correlation exists between paired means of many such pairs of samples. In such a case we may compute the difference between each pair of items in the two samples, and ascertain from these differences their mean $\overline{X}_D$, and their standard deviation σ_D or $\bar{\sigma}_D$. The value of $\sigma_{\overline{X}_D} = \dfrac{\bar{\sigma}_D}{\sqrt{N}}$ is then determined, and $\overline{X}_D$ is compared with $\sigma_{\overline{X}_D}$ to ascertain if $\overline{X}_D$ differs significantly from zero. Alternately, we may apply the test described in the text above. If either of these tests discredits the null hypothesis, we should not ignore its testimony.

The standard error of the difference is:

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{(1.84)^2 + (1.88)^2} = \sqrt{3\ 386 + 3.534,}$$
$$= \sqrt{6.920} = 2.63.$$

Having now the two values: (1) the observed difference between the two means $\bar{X}_1 - \bar{X}_2 = 110.52 - 109.48 = 1.04$, and (2) the standard error of the difference between the two means $= 2.63$, we are in a position to answer the following question:

If the true difference between the means is zero, what is the probability that $\bar{X}_1$ *might exceed* $\bar{X}_2$ *by 1.04 or more because of chance variations?*

The ratio of the observed difference to the standard error of the difference,

$$\frac{x}{\sigma} = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}_1 - \bar{X}_2}} = \frac{1.04}{2.63} = .395,$$

indicates the point on the curve beyond which all X values are as great or greater than the observed difference (see Chart 133). If we look up
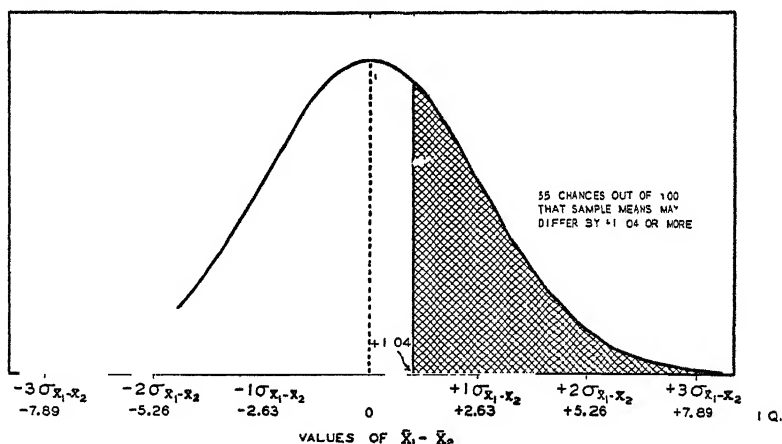
Chart 133.   Expected Distribution of Differences Between Sample Means of Intelligence Quotients of Left-Handed and Right-Handed Students and Chances of Obtaining a Difference of $+1.04$, When $\sigma_{\bar{X}_1 - \bar{X}_2} = 2.63$.

.395 in the table of areas of the normal curve (Appendix E) and subtract the result from .5000, we shall know the fraction of the area of the curve which is cross-hatched in Chart 133 and we shall have the answer to our question.   Looking up $\frac{x}{\sigma} = .395$ in Appendix E, we find .1536.   This means that 15 out of 100 sample differences would fall between $\bar{X}_1 - \bar{X}_2 = 0$ and $\bar{X}_1 - \bar{X}_2 = +1.04$, and consequently 50 — 15. or 35 out of 100,

would occur beyond $+1.04$. If the I.Q.'s of left-handed and right-handed students are actually identical, we should expect sample means of left-handed students to exceed those for right-handed students 50 times out of 100. It appears, then, that if the I.Q.'s of left-handed and right-handed students are actually identical, we might get differences between sample means as great as this and in favor of left-handed students 35 times out

TABLE 69

WEIGHT AT DEMOBILIZATION OF 1,821 SCOTCH* SOLDIERS SERVING IN THE UNITED STATES ARMY DURING THE WORLD WAR

Weight in pounds	f	d'	fd'	$f(d')^2$
100 and under 110	12	-4	$-$ 48	192
110 and under 120	79	-3	-237	711
120 and under 130	254	-2	-508	1,016
130 and under 140	436	-1	-436	436
140 and under 150	404	0	0	0
150 and under 160	308	$+1$	308	308
160 and under 170	175	$+2$	350	700
170 and under 180	89	$+3$	267	801
180 and under 190	37	$+4$	148	592
190 and under 200	19	$+5$	95	475
200 and over †	8	$+6$	48	288
Total .	1,821	.	$-$ 13	5,519

* Soldiers classified as Scotch were either (1) born in Scotland, or (2) had parents who were both born in Scotland, or (3) had three or four grandparents born in Scotland
† Mid-value taken as 205 pounds, giving results in agreement with those of the source
Source Charles B Davenport and Albert G Love *The Medical Department of the United States Army* ... in War, Vol XV, Part 1, pp 60 and 135, ... Office, Washington, 1921.

$$\overline{X} = 145 - \left(\frac{13}{1821}\right) 10 = 144.93 \text{ pounds.}$$

$$\sigma = 10 \sqrt{\frac{5519}{1821} - \left(\frac{13}{1821}\right)^2} = 17.41 \text{ pounds.}$$

$$\bar{\sigma} = 10 \sqrt{\frac{5519}{1820} - \frac{(13)^2}{1821 \cdot 1820}} = 17.41 \text{ pounds.}$$

of 100, owing to chance factors. This difference may have been due to chance and there is thus no definite evidence (from these data) that the I.Q. of the left-handed students is superior to that of right-handed students.

It will be recalled that, for the 746 soldiers of French extraction, $\overline{X}$ = 142.16 pounds and $\bar{\sigma}$ = 16.04 pounds. Table 69 shows a distribution of the weights of a group of 1,821 United States soldiers of Scotch extraction. The measurements were taken at demobilization in 1919, as were the others. From this table it is computed that $\overline{X}$ = 144.93 pounds and $\bar{\sigma}$ = 17.41 pounds. Is there a real difference in favor of the soldiers of

Scotch extraction, or is the difference so slight that it may be attributable to chance? If the difference is significant, we may regard the two groups as clearly different; if it is not significant, we conclude that they may have been drawn from the same population We shall proceed exactly as in the preceding example, computing $\sigma_{\bar{X}}$ for each series, then $\sigma_{\bar{X}_1 - \bar{X}_2}$, and finally comparing the observed difference of the means of the two series with the standard error of the difference. Summarizing:

Soldiers of Scotch extraction

$N_1 = 1{,}821$

$\overline{X}_1 = 144\,93$ pounds

$\tilde{\sigma}_1 = 17\,41$ pounds

$\sigma_{\bar{X}_1} = \dfrac{17\,41}{\sqrt{1{,}821}} = .408$ pounds

Soldiers of French extraction

$N_2 = 746$

$\overline{X}_2 = 142\,16$ pounds

$\tilde{\sigma}_2 = 16.04$ pounds

$\sigma_{\bar{X}_2} = \dfrac{16\,04}{\sqrt{746}} = .587$ pounds

$$\overline{X}_1 - \overline{X}_2 = 144.93 - 142\,16 = 2\,77 \text{ pounds}$$

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{(587)^2 + (408)^2} = 715 \text{ pounds}$$

6 CHANCES OUT OF 100 000 THAT SAMPLE MEANS MAY DIFFER BY +2 77 POUNDS OR MORE

+2.77

| $-3\sigma_{\bar{X}_1-\bar{X}_2}$ | $-2\sigma_{\bar{X}_1-\bar{X}_2}$ | $-1\sigma_{\bar{X}_1-\bar{X}_2}$ | 0 | $+1\sigma_{\bar{X}_1-\bar{X}_2}$ | $+2\sigma_{\bar{X}_1-\bar{X}_2}$ | $+3\sigma_{\bar{X}_1-\bar{X}_2}$ | $+4\sigma_{\bar{X}_1-\bar{X}_2}$ |
| -2 145 | -1.430 | - 715 | +.715 | +1 430 | +2,145 | +2.860 POUNDS |

VALUES OF $\bar{X}_1 - \bar{X}_2$
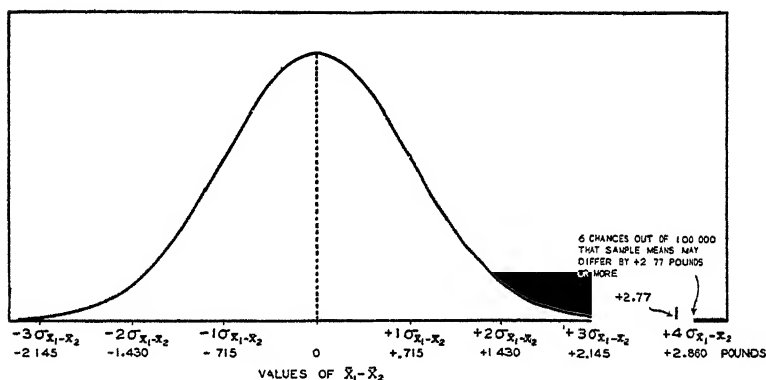
Chart 134. Expected Distribution of Differences Between Sample Means of Weight of American Soldiers of Scotch and French Extraction and Chances of Obtaining a Difference of +2.77 Pounds, When $\sigma_{\bar{X}_1 - \bar{X}_2} = .715$ Pounds.

The soldiers of Scotch extraction were 2.77 pounds heavier, on the average, than were those of French extraction. If the true difference between the mean weights of these two groups is zero, what are the chances that an observed difference of 2.77 in favor of the Scotch might occur, due to variations arising from sampling? The ratio of the observed difference to the standard error of the difference is

$$\frac{x}{\sigma} = \frac{\overline{X}_1 - \overline{X}_2}{\sigma_{\bar{X}_1 - \bar{X}_2}} = \frac{2.77}{.715} = 3.87.$$

From the table of areas of the normal curve, we find that $\dfrac{49{,}994}{100{,}000}$ of a

normal curve is between $\frac{x}{\sigma} = 0$ and $\frac{x}{\sigma} = 3.87$. Beyond the $\frac{x}{\sigma} = 3.87$ point, then, there are $50{,}000 - 49{,}994$ or 6 out of $100{,}000$ occurrences. This is shown in Chart 134 It is quite apparent that a difference of 2.77 pounds in favor of the Scotch soldiers could hardly occur fortuitously and there is thus a clearly significant difference between the two groups.

Procedure when $N_1 \neq N_2$. When testing the hypothesis that both samples have been drawn from the same population (i.e., that the true difference between the means is zero), we have used the formula

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2}}$$
$$= \sqrt{\frac{\bar{\sigma}^2_1}{N_1} + \frac{\bar{\sigma}^2_2}{N_2}}.$$

It is, however, more accurate to utilize all available information to make one estimate of the variance of the parent population from the variances of the two samples taken together, and to substitute that in place of the two separate estimates in the above formula. The formula then becomes

$$\sigma'_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\bar{\sigma}^2_{1+2}}{N_1} + \frac{\bar{\sigma}^2_{1+2}}{N_2}},$$

in which $\bar{\sigma}^2_{1+2}$ is the average of the two estimates of population variance computed by the expression

$$\bar{\sigma}^2_{1+2} = \frac{\Sigma x_1^2 + \Sigma x_2^2}{(N_1 - 1) + (N_2 - 1)}.$$

When $N_1 = N_2$, the results are identical regardless of whether two sep arate estimates or one combined estimate of variance is used; but when $N_1 \neq N_2$, in order to obtain maximum accuracy $\bar{\sigma}^2_1$ and $\bar{\sigma}^2_2$ must be weighted by their respective degrees of freedom when we are averaging them to obtain the final estimate of $\bar{\sigma}^2$. The result is the more complicated expression[12]

$$\sigma'_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{(N_1 + N_2)(\Sigma x_1^2 + \Sigma x_2^2)}{N_1 N_2 [(N_1 - 1) + (N_2 - 1)]}}.$$

While this expression introduces an element of increased rigor into the method, it is not apt to alter the conclusion appreciably when the N's are large, but may be important when the N's are small and unequal. Very rarely do we have to compare series with N's as greatly different as for the soldiers of French and Scotch extraction. The value of Σx_1^2 (for the Scotch) may be obtained from

$$\Sigma x_1^2 = \Sigma X_1^2 - \frac{(\Sigma X_1)^2}{N},$$

[12] See Appendix B, section XII–5.

or, since we are dealing with a frequency distribution, from

$$\Sigma x_1^2 = i^2 \left[\Sigma f(d')^2 - \frac{(\Sigma fd')^2}{N} \right];$$

and similarly for Σx_2^2 for the French. [The value of Σx_1^2 may also be obtained from $N_1 \sigma_1^2$ or $(N_1 - 1)\hat{\sigma}_1^2$, and similarly for Σx_2^2.] Referring to Table 69, we find the values for the Scotch group are:

$$\Sigma x_1^2 = 100 \left[5,519 - \frac{(13)^2}{1,821} \right]$$

$$= 100 (5,519 - .09281)$$

$$= 551,890.72.$$

Taking the necessary totals from Table 68, we find the values for the French group are:

$$\Sigma x_2^2 = 100 \left[1,978 - \frac{(212)^2}{746} \right]$$

$$= 100 (1,978 - 60.24665)$$

$$= 191,775.34.$$

The value of $\sigma'_{\bar{x}_1 - \bar{x}_2}$ may now be determined:

$$\sigma'_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{(1,821 + 746)(551,890.72 + 191,775.34)}{(1,821)(746)(1,820 + 745)}}$$

$$= \sqrt{\frac{(2,567)(743,666.06)}{(1,358,466)(2,565)}}$$

$$= \sqrt{\frac{1,908,990,776.02}{3,484,465,290}}$$

$$= \sqrt{.547858}$$

$$= .740.$$

Comparing the observed difference between the means to this value gives the ratio

$$\frac{x}{\sigma} = \frac{\bar{X}_1 - \bar{X}_2}{\sigma'_{\bar{x}_1 - \bar{x}_2}} = \frac{2.77}{.740} = 3.74.$$

Referring to the table of areas of the normal curve, it appears that $\frac{49,991}{100,000}$ of the area of the curve occurs between $\frac{x}{\sigma} = 0$ and $\frac{x}{\sigma} = +3.74$. Beyond $\frac{x}{\sigma} = +3.74$, therefore, we should find $\frac{9}{100,000}$ of the area. Since there is about one chance out of 10,000 that the observed difference in favor of the

Scotch might occur fortuitously, it follows that the difference is clearly significant. It may be observed that the probability obtained by this method and the probability obtained by the preceding method are in rather close agreement even though the N's are respectively 746 and 1,821.

Reliability of the mean of a stratified sample. What has been said up to this point has dealt with the reliability of means based on *random* samples. Sometimes, however, we can obtain even more reliable results by using a *stratified* sample. Thus, when the population can be divided into pertinent categories or *strata*, we may select samples at random from each of these strata This procedure is especially apropos when the population is heterogeneous and is composed of a number of strata each relatively homogeneous. A stratified sample is usually so selected that the number included in the sample from each stratum is proportional to the numerical importance of that stratum in the population.

When information concerning the population is available, the population variance, used in computing the standard error of the mean, is based on the deviations of the items within each stratum from the mean of that stratum, rather than from the mean of the entire population. Referring to this as $\sigma_{P'}^2$ to distinguish it from σ_P^2, computed in relation to the mean of the entire population, we have

$$\sigma_{P'}^2 = \frac{\sum\limits_{1}^{m}\sum\limits_{a}^{P_S}(X - \overline{X}_S)^2}{\sum\limits_{1}^{m}P_S},$$

where P_S is the number of items in a stratum, $\overline{X}_S$ is the mean of a stratum, $\sum\limits_{a}^{P_S}$ indicates a summation from item a to item P_S in a stratum, and $\sum\limits_{1}^{m}$ indicates a summation from stratum 1 to stratum m.

Now if $(X - \overline{X}_S)^2$ is summed for a particular stratum the value obtained is $P_S\sigma_S^2$, where σ_S^2 is the variance of the stratum. Also, $\Sigma P_S = P$, the number of items in the population, and we may write

$$\sigma_{P'}^2 = \frac{\sum\limits_{1}^{m}P_S\sigma_S^2}{P}.$$

Therefore

$$\sigma_{\overline{X} \text{ of a stratified sample}}^2 = \frac{\sigma_{P'}^2}{N} = \frac{\sum\limits_{1}^{m}P_S\sigma_S^2}{P} \div N,$$

where N is the number of items in the entire sample. In Appendix B, Section XII–6, it is shown that

$$\sigma_{\overline{X} \text{ of a stratified sample}}^2 = \frac{\sigma_P^2}{N} - \frac{\sigma_{\text{of strata means}}^2}{N},$$

where

$$\sigma^2_{\text{of strata means}} = \frac{\sum\limits_1^m P_S(\bar{X}_S - \bar{X}_P)^2}{P}.$$

Either of these expressions is satisfactory (since they are equivalents) when population data are available. However, when we have to base our computations on sample data we refer to the procedure used for two samples on page 322 and merely extend the expression to cover several strata as if they were several samples   Using the symbol $\bar{\sigma}^2_{1\cdots m}$ to indicate the estimate of σ^2_P based on several samples, we have

$$\bar{\sigma}^2_{1\cdots m} = \frac{\sum x_1^2 + \sum x_2^2 + \cdots + \sum x_m^2}{(N_1 - 1) + (N_2 - 1) + \cdots + (N_m - 1)}$$

$$= \frac{n_1\bar{\sigma}_1^2 + n_2\bar{\sigma}_2^2 + \cdots + n_m\bar{\sigma}_m^2}{n_1 + n_2 + \cdots + n_m},$$

and the standard error of the mean of a stratified sample becomes

$$\sigma_{\bar{X} \text{ of a stratified sample}} = \frac{\bar{\sigma}_{1 \cdot m}}{\sqrt{N}},$$

if the population is large in relation to the sample.

Reliability of Sample Means, Small Samples

The t distribution. In the preceding discussion it was assumed that the sampling distribution of $\dfrac{\bar{X} - \bar{X}_P}{\sigma_{\bar{X}}}$ is normal, which is essentially true when N is large but not when N is small. Similarly, the distribution of $\dfrac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{x}_1 - \bar{x}_2}}$ is effectively normal when N is large but not when N is small. In each instance the lack of normality arises out of the fact that, in computing the value of $\sigma_{\bar{X}}$, we use $\bar{\sigma}$ (the estimate of the standard deviation of the population) in place of the true value of the standard deviation of the population σ_P.[13]   When N is small, the sampling distribution follows the t curve, which is more widely dispersed than the normal curve (see Chart 135) and becomes more so in inverse relation to the degrees of freedom present.   As may be seen from the chart, there is a greater proportion of the t curve beyond any given deviation from the mean than there is in the case of the normal curve.   A table (called the t table) has been devised, which gives for various degrees of freedom (n) the probability that observed values differing from zero (either positively or negatively) may occur owing to chance.   If a quantity is distributed normally in terms

---

[13] For a more detailed discussion, see R. A. Fisher, *Statistical Methods for Research Workers*, pp. 124–125, Oliver and Boyd, Edinburgh, 1938 (7th Edition).
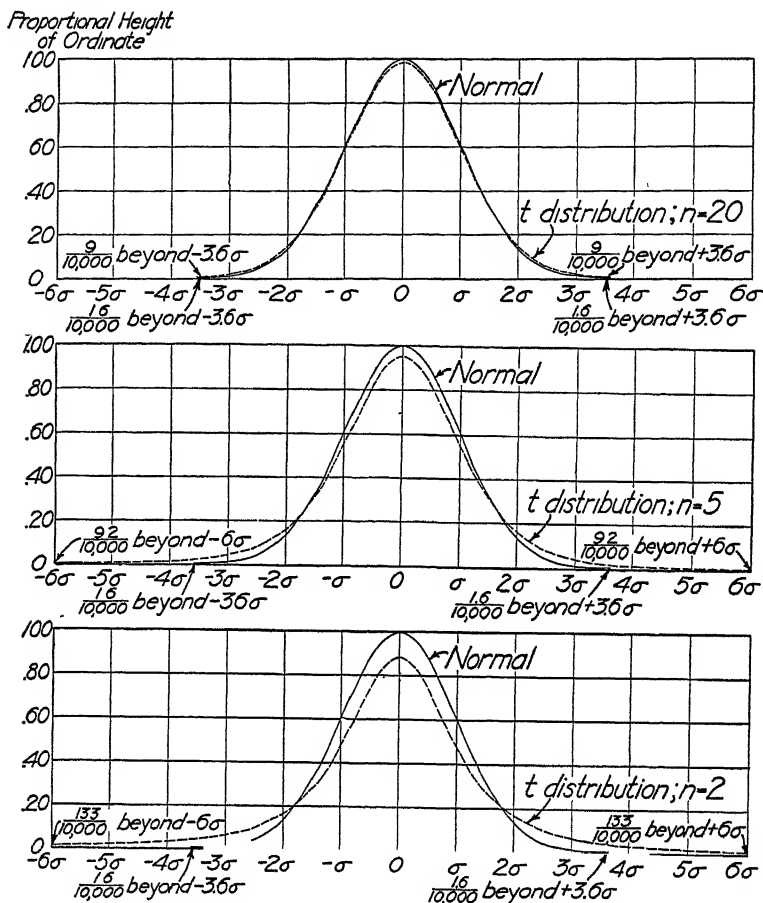
Chart 135. Comparison of the Normal Distribution and the t Distribution when $n = 20$, $n = 5$, and $n = 2$. The ordinates of the t distribution are obtained by the

expression $Y_c = \sqrt{\dfrac{2}{n}} \; \dfrac{\left(\dfrac{n-1}{2}\right)!}{\left(\dfrac{n-2}{2}\right)!} \; \dfrac{1}{\left(1+\dfrac{t^2}{n}\right)^{\frac{n+1}{2}}}$. This gives a maximum ordinate of

1 0000, comparable to the expression $Y_c = e^{\frac{-x^2}{2\sigma^2}}$ for the normal curve. The symbol t^2 in the expression for the t distribution is the same as $\dfrac{x^2}{\sigma^2}$ in the expression for the normal

curve. The computation of $\dfrac{\left(\dfrac{n-1}{2}\right)!}{\left(\dfrac{n-2}{2}\right)!}$ may be clarified by an illustration. If $n = 11$,

the numerator is 5!, while the denominator is 4.5!. The value of 4.5! is given by 4.5 $\times$ 3 5 $\times$ 2.5 $\times$ 1 5 $\times$ .5 $\times$ $\sqrt{\pi}$.

of its *actual* standard error around a mean of zero, t is the ratio of that quantity to an *estimate* of its standard error. Thus, while

$$\frac{\overline{X} - \overline{X}_P}{\dfrac{\sigma_P}{\sqrt{N}}}$$

is distributed normally,

$$\frac{\overline{X} - \overline{X}_P}{\dfrac{\bar{\sigma}}{\sqrt{N}}} = t$$

is distributed according to the t distribution. Since $\bar{\sigma}$ differs little from σ_P when N is large, this is an important consideration only when N is small.

An examination of this table (shown in Appendix F) or of Chart 135 will reveal the wider range of the t distribution in comparison with the normal distribution. In a normal distribution, 99 per cent of a series of sample means would be within $\doteq 2.6\sigma_{\overline{x}}$ of the population mean; in a t distribution, with $n = 10$ ($N = 11$), 99 per cent would be within $\doteq 3.20\sigma_{\overline{x}}$, a somewhat wider range. This table should be used whenever the value of N is small; that is to say, when the degrees of freedom are 30 or fewer, although the difference between this table and the table of areas of the normal curve does not become appreciable until N becomes less than 20. This may be seen by referring to the t table in Appendix F. The last line of the t table, which shows the values of t when $n = \infty$, gives the same ratios as are to be had for $\dfrac{x}{\sigma}$ from the table of areas of the normal curve.

Reliability of a sample mean when N is small. Tests have been made of the breaking strength of ten pieces of hard-drawn copper wire as shown in Table 70. The mean of this sample is 575.2 pounds, and it is desired to ascertain the reliability of this sample mean. As shown below the table, $\bar{\sigma} = 8.70$ pounds and $\sigma_{\overline{x}} = 2.75$ pounds. The value of σ is 8.26 pounds, and it is to be noted that this value is appreciably smaller than $\bar{\sigma}$ because N is small.

In studying the reliability of the mean from this small sample, let us first assume that we know the mean of an entire population and test the divergence of our sample mean from this population mean. Suppose that the population mean is known to be 577.0. Is the value of the sample mean (575.2) sufficiently different to invalidate the hypothesis that the sample was drawn from this population? The deviation of the sample mean is -1.8 pounds, which is .65 times the standard error of the mean. By referring to the t table of Appendix F, for $n = 9$ ($n = N - 1$ since 1 degree of freedom was lost when $\overline{X}$ was computed) and $t = .65$, we find, by interpolating, that there are about 53 chances out of 100 that a mean

from such a sample may differ ± 1.8 pounds or more from the population mean. There are about 26 chances out of 100 that such a sample mean might fall *below* the population mean by 1.8 pounds or more. There is,

<div align="center">

TABLE 70

BREAKING STRENGTH OF 10 SPECIMENS OF .104-INCH
DIAMETER HARD-DRAWN COPPER WIRE

</div>

Specimen	Breaking strength in pounds X	X^2
1	578	334,084
2	572	327,184
3	570	324,900
4	568	322,624
5	572	327,184
6	570	324,900
7	570	324,900
8	572	327,184
9	596	355,216
10	584	341,056
Total	5,752	3,309,232

Source. American Society for Testing Materials, *Supplements to 1933 A S T M Manual on Presentation of Data,* "Supplement A—Presenting Plus and ˮ ᵘ ᵗ ᵗᵗ ᶦ ' ᵗ ᵗ ˡ' ᵗ ᵗ ̧ ᵗᵗ ᵗᵗ ᵗ ᵗ ᵗ ᵗ ᵗ ᵗ ᵗᵗ ᵗ ᵗ ᵗ ᵗ ᵗ ᵗ ᵗᵗ ᵗ ᵗ ᵗ ᵗ ᵗ ᵗ ᵗᵗ ᵗ ᵗ ᵗ ᵗ ᵗ ᵗᵗ ᵗ ty of an Observed Average" p 1, reprinted ᵗ ᵗ ᵗᵗ ᵗ ᵗᵗ ᵗ ᵗ ᵗᵗ ᵗ American ᵗ ᵗ ᵗ ᵗ ᵗ American Society ᵗ ᵗ ᵗ Materials, V ᵗ ᵗ ᵗ ᵗ ᵗ ᵗ ᵗ ᵗᵗ ᵗ 1935

$$\bar{X} = \frac{5752}{10} = 575.2 \text{ pounds.}$$

$$\sigma = \sqrt{\frac{3,309,232}{10} - \left(\frac{5752}{10}\right)^2} = 8.26 \text{ pounds.}$$

$$\tilde{\sigma} = \sqrt{\frac{3,309,232}{9} - \frac{(5752)^2}{10 \cdot 9}} = 8.70 \text{ pounds.}$$

$$\sigma_{\bar{X}} = \frac{8.70}{\sqrt{10}} = 2.75 \text{ pounds.}$$

consequently, no clear evidence of a significant divergence of the sample mean from the population mean, and the sample may well have been drawn from this population.

Now let us assume, as is more frequently the case, that we do not know the value of the population mean but wish to form some idea of the range within which its value may occur. Our computed values are, as before, $\bar{X} = 575.2$ pounds and $\sigma_{\bar{X}} = 2.75$ pounds; also, $N = 10$ and $n = 9$. If we were dealing with a large sample, the sample means would vary $\pm 2\sigma_{\bar{X}}$ or more from the population mean in about 5 samples out of 100 (see area table, Appendix E). However, we are considering a small sample with 9

degrees of freedom.   Referring to the t table (Appendix F), opposite $n = 9$ we observe that the sample means would vary $\pm 2.3\sigma_{\bar{X}}$ or more from the population mean in about 5 samples out of 100   Assuming that the ".05 level of significance" is satisfactory as a criterion, we find that, if the observed mean is below the population mean, the population mean might be $575.2 + (2.3)(2.75) = 581\ 5$ pounds; while, if the observed mean is above the population mean, the population mean might be $575.2 - (2.3)(2.75) = 568.9$ pounds.   We conclude, then, that it is likely that the population mean falls between 568.9 pounds and 581.5 pounds, since the fiducial probability is 95 out of 100 that the population mean falls between these limits. If a more strict criterion is desired, we may consider the range which will include (say) 98 per cent of the possible sample values.   Entering the t table at $n = 9$, we find that 98 per cent of the sample means would vary within $\pm 2.8\sigma_{\bar{X}}$ ($\pm 2.8 \times 2.75 = \pm 7.7$ pounds) of the population mean. It is even more likely, then, that the population mean is between 567.5 and 582.9 pounds, as the fiducial probability is 98 out of 100 that the population mean occurs within this range

TABLE 71

STRENGTH OF LEAD IN TWO NUMBER 2 PENCILS MANUFACTURED BY "COMPANY E"

Pencil (a)   Pencil (b)

Test	Strength in kilograms X_1	X_1^2	Test	Strength in kilograms X_2	X_2^2
1	1.62	2.6244	1	1.78	3.1684
2	1.74	3 0276	2	1.48	2 1904
3	1.68	2.8224	3	1.72	2.9584
4	1.50	2.2500	4	1.62	2.6244
Total	6.54	10 7244	Total	6.60	10 9416

Source: From tests conducted in 1934 by the Eagle Pencil Company

$$\bar{X}_1 = \frac{6.54}{4} = 1.635 \text{ kilograms}$$

$$\bar{X}_2 = \frac{6.60}{4} = 1.650 \text{ kilograms.}$$

$$\bar{\sigma}_1 = \sqrt{\frac{10\ 7244}{3} - \frac{(6.54)^2}{4 \cdot 3}}$$

$$\bar{\sigma}_2 = \sqrt{\frac{10\ 9416}{3} - \frac{(6.60)^2}{4 \cdot 3}}$$

$$= 1015 \text{ kilograms.}$$

$$= .1311 \text{ kilograms.}$$

$$\sigma_{\bar{X}_1} = \frac{.1015}{\sqrt{4}} = .0507 \text{ kilograms.}$$

$$\sigma_{\bar{X}_2} = \frac{.1311}{\sqrt{4}} = .0656 \text{ kilograms.}$$

Significance of the difference between two means when N is small. Table 71 shows data of the strength of the points of two Number 2 pencils manufactured by a company designated as "Company E."   Let us ascer-

tain if the difference in the mean strength of pencils (a) and (b) made by Company E is significant. From Table 71 it is computed that $\overline{X}_1 = 1.635$ and $\overline{X}_2 = 1.650$, while $\sigma^2_{\overline{X}_1} = .0026$ and $\sigma^2_{\overline{X}_2} = .0043$ kilograms. The standard error of the difference between the means is

$$\sigma_{\overline{X}_1 - \overline{X}_2} = \sqrt{.0026 + .0043} = .0831 \text{ kilograms},$$

and

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sigma_{\overline{X}_1 - \overline{X}_2}} = \frac{1.635 - 1.650}{.0831} = \frac{.015}{.0831} = .18.$$

This is the t of the t table (Appendix F) which we consider in conjunction with n, the degrees of freedom present. In this instance, since the mean was computed for each series, there are three degrees of freedom present in each series: $n_1 = 3$, $n_2 = 3$; $n = n_1 + n_2 = 6$. When $n = 6$, $t = .131$ for $P = .9$, and $t = .265$ for $P = .8$. Therefore, when $t = .18$, P is about .86, indicating that a difference of $\pm.015$ kilograms might occur about 86 times in 100. There is thus no evidence of a significant difference in strength between the points of these two pencils.

Significance of the difference between two means when $N_1 \neq N_2$ and when both N's are small. From archaeological excavations conducted at a certain site, 16 lower first molars were recovered.[14] These showed a mean length $\overline{X}_1$ of 13.57 millimeters and σ_1 of .72 millimeters. From a nearby site, 9 lower first molars were taken with $\overline{X}_2 = 13.06$ and $\sigma_2 = .62$ millimeters. Is there a significant difference in the mean length of these two groups of lower first molars? When $N_1 \neq N_2$, we pool the variances[15] of the two series by use of the expression shown on page 322.

$$\sigma'_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{(N_1 + N_2)(\Sigma x_1^2 + \Sigma x_2^2)}{N_1 N_2[(N_1 - 1) + (N_2 - 1)]}}$$

$$= \sqrt{\frac{(16 + 9)(8.2944 + 3.4596)}{16 \cdot 9 \,(15 + 8)}} = .298.$$

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sigma'_{\overline{X}_1 - \overline{X}_2}} = \frac{.51}{.298} = 1.71.$$

The first set of data contributes 15 degrees of freedom, while the second contributes 8, and $n = 15 + 8 = 23$. Referring to the t table, for $t = 1.71$

---

[14] Based upon illustrative figures used in a lecture by Professor Egon Pearson at Columbia University in 1931.

[15] We obtain Σx_1^2 from σ_1 as follows:

$$\sigma_1 = \sqrt{\frac{\Sigma x_1^2}{N_1}}; \qquad \sigma_1^2 = \frac{\Sigma x_1^2}{N_1}; \qquad \Sigma x_1^2 = N_1 \sigma_1^2.$$

Therefore $\Sigma x_1^2 = 16(.72)^2 = 8.2944$. Similarly, for the second set of data, $\Sigma x_2^2 = 9(.62)^2 = 3.4596$.

and $n = 23$, we find that there is about 1 chance in 10 that such an observed difference ($\pm.51$ millimeters) might occur by chance and we conclude that there is no clear evidence that the difference is significant.

If, instead of using the t table, we had referred to the table of areas of the normal curve, we would have found that for $\frac{x}{\sigma} = 1.71$ the chances of obtaining a difference of $\pm.51$ were 872 out of 10,000, or 8.7 out of 100. While the difference in the P values is not great when $n = 23$, it will be found to be progressively greater as n becomes smaller.

In the preceding paragraphs we have considered the significance of differences between single pairs of means taken at random. If instead of a single pair of means obtained from two samples we have a hundred sample differences the true value of which is zero, we would expect about five of the differences to be beyond the range of twice the standard error; one would be expected to be beyond the range of 2.6 times the standard error. Such a great difference might erroneously be reported as significant. It is not within the scope of this text to discuss this problem. Tippett[16] suggests a procedure whereby in testing the significance of the greatest difference between two samples out of a group of samples the usual procedure is slightly modified. The method of analysis of variance, discussed in the following chapter, may also be applied.

Selected References

R. A. Fisher. *Statistical Methods for Research Workers* (Seventh Edition), pages 120–133; Oliver and Boyd, Edinburgh, 1938.

F. C Mills: *Statistical Methods Applied to Economics and Business* (Revised Edition), pages 452–473, 481–483; Henry Holt and Co., New York, 1938. Fiducial limits are discussed on pages 465–466.

P. R. Rider. *An Introduction to Modern Statistical Methods*, pages 77–81, 88–93; John Wiley and Sons, New York, 1939. Includes brief discussion of fiducial limits.

W. A. Shewhart· *Economic Control of Quality of Manufactured Product*, Chapters XIII, XIV, XXIII, XXIV; D. Van Nostrand Co., New York, 1931. Discussion of sampling fluctuations in Chapters XIII, XIV.

G. W. Snedecor: *Statistical Methods Applied to Experiments in Agriculture and Biology*, page 30, Chapters 3, 4; Collegiate Press, Ames, Iowa, 1937. A statement concerning degrees of freedom appears on page 30. Fiducial limits are treated on pages 59–61.

L. H. C. Tippett: *The Methods of Statistics* (Second Edition), pages 62–82, 110–116; Williams and Norgate, London, 1937. The sampling distribution of the mean is discussed on pages 64–66. The acceptance and rejection of hypotheses are treated on pages 74–78.

A. E. Waugh: *Elements of Statistical Method*, pages 131–140, 148–157; McGraw Hill Book Co., New York, 1938.

---

[16] See L H. C. Tippett, *The Methods of Statistics,* pp. 78–80, Williams and Norgate. London, 1937 (2nd Edition).

CHAPTER XIII
RELIABILITY AND SIGNIFICANCE OF STATISTICAL MEASURES
PERCENTAGES, STANDARD DEVIATIONS, VARIANCES, AND THE CRITERION OF LIKELIHOOD

Reliability of Sample Percentages

Standard error of a sample percentage. If a percentage (or a proportion) has been computed from a sample drawn at random from a larger population, it is subject to sampling variations, as is any other statistical measure. The standard error of a percentage σ_p is given by the expression

$$\sigma_p = \sqrt{\frac{pq}{N}} \text{ or } \sqrt{\frac{p - p^2}{N}},$$

where p is the proportion in the population expressed as a decimal, $q = 1 - p$, and N is the number of items in the sample.

Reliability of a sample percentage. If the proportions of white and black marbles in a large assortment are equal, we have p (proportion white) $= .50$ and q (proportion black) $= .50$. Assuming a normal distribution of sample proportions about the proportion in the population we find that if random samples of 100 marbles are drawn we should expect about 68 out of 100 of such samples to show proportions of white marbles varying within $\pm \sigma_p$ of $.50$. Since

$$\sigma_p = \sqrt{\frac{(.50)(.50)}{100}} = \sqrt{\frac{.25}{100}}$$

$$= \sqrt{.0025} = .05,$$

the expected range is from 45 to 55 per cent. Likewise we should expect about 95 out of 100 samples to show from 40 to 60 per cent of white marbles, and about 99 out of 100 to show from 35 to 65 per cent of white marbles.

In a group of 40 first cousins there were found to be 22 males and 18 females. Let us ascertain if the observed proportions are inconsistent with the hypothesis that the sexes should be in equal proportions. Letting

p represent the proportion of males and q the proportion of females in the population, we have:

$$\sigma_p = \sqrt{\frac{(.50)(.50)}{40}} = \sqrt{\frac{.25}{40}}$$
$$= \sqrt{.00625} = .079.$$

The observed proportion of males p_S was $\frac{22}{40}$, or 55 per cent, while the proportion in the population was 50 per cent. This gives

$$\frac{x}{\sigma} = \frac{p_S - p}{\sigma_p} = \frac{.55 - .50}{.079} = \frac{.05}{.079} = .63.$$

Looking up this ratio in the table of areas of the normal curve shows that a divergence of 5 per cent or more in favor of males might occur in about 26 samples in 100, and that a divergence of 5 per cent or more in either direction might be expected to occur in about 52 samples in 100. We therefore conclude that the observed proportions do not show a significant divergence from the population values.

In October 1911 the Chicago, Milwaukee, St. Paul & Pacific Railway installed a number of ties near Hartford, Wisconsin, for the purpose of testing various kinds of wood and various materials and processes for preserving the wood. One lot of 50 red oak ties was preserved by means of creosote applied by the "full cell" process. In June 1934, after about $22\frac{3}{4}$ years of service, 22 ties, or 44 per cent, were still in good condition.[1] How far may this value be from the population percentage? Since the values of p and q in the population are not known, we obtain a rough approximation by substituting the proportions found in the sample. Thus

$$\sigma_p = \sqrt{\frac{(.44)(.56)}{50}} = \sqrt{\frac{2464}{50}} = .070, \text{ or } 7.0 \text{ per cent.}$$

The chances are therefore about 95 out of 100 that the observed percentage is within $\pm 1.96 \times .07$ (or 13.7 per cent) of the population percentage. Further, the chances are 99 out of 100 that the observed percentage lies within the range of $\pm 2.58\sigma_p$ (or 18.1 per cent) of the population percentage. The reliability of this percentage is thus quite low. In terms of fiducial probabilities we might say that there is a 95 per cent fiducial probability that the population percentage lies between about 30 and 58 per cent, and a 99 per cent fiducial probability that the population percentage lies between about 26 and 62 per cent. A somewhat more satisfactory determination of the fiducial limits is given on pages 335–337.

Reliability of percentages and the χ^2 test. Let us consider another procedure for evaluating the reliability of ratios. In the group of 40 first

[1] The data are from *Proceedings of the American Wood Preservers Association*, 1935, pp. 133–134.

cousins mentioned before, there were found to be 22 males and 18 females. Again we shall ascertain if this distribution is inconsistent with the hypothesis that the sex ratio is 1 : 1. We shall proceed to make the χ^2 test exactly as was done for testing the fit of a normal curve in an earlier chapter. The results are as follows:

Sex	Observed f	Expected, ratio 1 . 1 f_c	$f - f_c$	$(f - f_c)^2$	$\dfrac{(f - f_c)^2}{f_c}$
Male	22	20	+2	4	.20
Female	18	20	−2	4	.20
Total .. .	40	40	. .	.	40

The value of χ^2 is .40 and there are two categories, male and female. The two sets of data were brought into agreement in respect to totals and thus one degree of freedom was lost. For $n = 1$ and $\chi^2 = .40$, we find P (see Appendix I) is slightly more than .50. The proportion of the distribution of χ^2 beyond $\chi^2 = .40$, when $n = 1$, may be visualized by referring to the chart accompanying Appendix I. The test indicates that the difference from expectation could well have arisen from chance and therefore does not indicate a significant digression from the expected. This conclusion is the same as that arrived at earlier.[2]

The value of χ^2 may also be obtained from the expression[3]

$$\chi^2 = \frac{\left(a - \dfrac{p}{q}b\right)^2}{\dfrac{p}{q}N},$$

where: $a =$ the number of times the first factor occurred;

$b =$ the number of times the second factor occurred;

$p =$ the expected or hypothetical probability associated with the first factor;

$q =$ the expected or hypothetical probability associated with the second factor;

$N = a + b$.

---

[2] Similar results will also be obtained by using the standard error of the number of occurrences, $\sigma_a = \sqrt{Npq}$, computing $\dfrac{a - pN}{\sigma_a}$ (where p is the proportion in the population and a is the actual number of occurrences in the sample corresponding to p), and referring to the table of areas of the normal curve.

[3] This expression may also be written

$$\chi^2 = \frac{(qa - pb)^2}{Npq}.$$

See Appendix B, section XIII-1, for a development of these two expressions from the usual expression for χ^2.

For the data of forty first cousins,

$$\chi^2 = \frac{\left(22 - \dfrac{.50}{.50}\,18\right)^2}{\dfrac{.50}{.50}\,40} = \frac{4^2}{40} = .40,$$

the same as was found before.

Upon the basis of past experience the fatality rate from typhoid fever for a certain community was found to be 14.2 per cent (that is, reported deaths from typhoid fever ÷ reported cases of typhoid fever = .1420). A survey was made of certain congested areas. The homes studied were selected as nearly as possible at random and a fatality ratio of 30.0 (36 deaths) was found for 120 cases of typhoid fever. Does this represent a significant departure from the population value 14.2? We shall compute χ^2 and determine the probability that such a value of χ^2 might arise by chance.

$$\chi^2 = \frac{\left(a - \dfrac{p}{q}\,b\right)^2}{\dfrac{p}{q}\,N} = \frac{\left(36 - \dfrac{.142}{.858}\,84\right)^2}{\dfrac{.142}{.858}\,120} = 24.59.$$

As before, there are two categories (that is, patients may survive or die), one degree of freedom has been lost since the observed and expected data have been brought into agreement with respect to totals, and $n = 1$. Such a value of χ' is far beyond the .001 level and we conclude that the difference is clearly significant.

We may also use the χ^2 expression just given to ascertain the fiducial limits of p from sample values of a, b, and N. Considering the data of 50 red oak ties creosoted by the "full cell" process which were given previously in this chapter, it was found that after a certain period of service 22 (or 44 per cent) of the ties were still in good condition. Thus $a = 22$, $b = 28$, $N = 50$. Let us first determine the 95 per cent fiducial limits for p. Referring to the first use of χ^2 in this chapter (for the determination of χ^2 for the sex distribution of first cousins), it will be apparent that whether the observed frequencies *exceed* or *fall below* the expected frequencies does not matter; either divergence makes χ^2 large and P small. Thus the 95 per cent fiducial level is given by taking the value of χ^2 at $P = .05$. (When $n = 1$ and $P = .05$, $\chi^2 = 3.841$.) The computations are

$$\chi^2 = \frac{\left(a - \dfrac{p}{q}\,b\right)^2}{\dfrac{p}{q}\,N}$$

$$3.841 = \frac{\left(22 - \frac{p}{q}\,28\right)^2}{\frac{p}{q}\,50}$$

$$192.050\,\frac{p}{q} = 484 - 1232\,\frac{p}{q} + 784\left(\frac{p}{q}\right)^2$$

$$484 - 1424.050\frac{p}{q} + 784\left(\frac{p}{q}\right)^2 = 0.$$

For a quadratic of the form $a + bX + cX^2 = 0$, when a, b, and c are given, we may solve for X by use of the expression

$$X = \frac{-b \pm \sqrt{b^2 - 4ac}}{2c}.$$

Thus

$$\frac{p}{q} = \frac{1424.050 \pm \sqrt{(1424.050)^2 - 4(784)(484)}}{2(784)}$$

$$= \frac{1424.050 \pm 714.029}{1568}$$

$$= \frac{2138.079}{1568} \text{ and } \frac{710.021}{1568}.$$

For the first of these fractions,

$$\frac{p}{q} = \frac{2138.079}{1568},$$

$$\frac{p}{p+q} = \frac{2138.079}{2138.079 + 1568},$$

$$\frac{p}{1} = .577 \text{ (or 57.7 per cent).}$$

Solving the second in similar fashion,

$$p = .312 \text{ (or 31.2 per cent).}$$

From the above we conclude that there is a 95 per cent fiducial probability that the population value of p (proportion of ties surviving) lies between 31.2 and 57.7 per cent.

If we desire to ascertain the 99 per cent fiducial limits of p, we use $\chi^2 = 6.635$, which gives

$$6.635 = \frac{\left(22 - \frac{p}{q}\,28\right)^2}{\frac{p}{q}\,50}$$

$$784 \left(\frac{p}{q}\right)^2 - 1563.750 \frac{p}{q} + 484 = 0$$

$$\frac{p}{q} = \frac{1563.750 \doteq 963.063}{1568}$$

$$p = .617 \text{ and } .277.$$

These results tell us that there is a 99 per cent fiducial probability that the population value of p falls between 27.7 and 61.7 per cent. These are non-equidistant limits* about the sample percentage in contrast to the rougher equidistant limits on page 333.

Significance of difference between percentages. At the same time the 50 red oak ties preserved with creosote by the "full cell" process were laid, another group of 50 red oak ties was installed. The second lot, however, was creosote-impregnated by the "Rueping" process. Of this lot, 18 ties, or 36 per cent, were still in service in June 1934. Assuming that both lots were subjected to identical conditions otherwise (and this appears to be true), is the difference between the percentages significant? For the "full cell" processed ties, p_1 was .44, or 44 per cent, and σ_{p_1} was found to be .070. For the "Rueping" processed ties,

$$\sigma_{p_2} = \sqrt{\frac{(.36)(.64)}{50}} = \sqrt{\frac{.2304}{50}} = .068, \text{ or } 6.8 \text{ per cent.}$$

The standard error of the difference between the two percentages is

$$\sigma_{p_1 - p_2} = \sqrt{\sigma_{p_1}^2 + \sigma_{p_2}^2}$$

$$= \sqrt{.070^2 + .068^2}$$

$$= .098, \text{ or } 9.8 \text{ per cent.}$$

The observed difference between the two percentages is $44 - 36 = 8$.

$$\frac{x}{\sigma} = \frac{p_1 - p_2}{\sigma_{p_1 - p_2}} = \frac{8}{9.8} = .82.$$

This ratio is so low that it is clear that the advantage of 8 per cent in favor of the "full cell" process may have been due to chance.

During a three-year period, experiments with two types of lighting systems were conducted in an elementary school. Room 1 had two 150-watt manually controlled lights, which were turned on and off by teacher or pupils as needed. Room 2 had four 300-watt indirect lights, controlled by automatic relays which turned the lights on when additional illumination was needed and off when no longer required. The pupils were sixth grade students. All were given the "Otis Self-Administered Test of Mental Ability" and the "Standard Achievement Test." They were divided

---

* A more accurate method is given by Clopper and Pearson in *Biometrica*, Vol. XXVI, pp. 404–413

equally between Room 1 and Room 2 according to the results of these tests. The two classrooms were identical second floor rooms with north light. Two teachers employing the same teaching methods were assigned to departmental work with the two classes, each teacher teaching certain subjects in both sections. During the three-year period there were 115 pupils in Room 1 and 112 pupils in Room 2. In Room 1 there were 29 failures; in Room 2 there were 9 failures. The percentages failing were 25.2 per cent for Room 1, and 8.0 per cent for Room 2.[4] Is there a significant difference between the two percentages?

$$\text{For Room 1:} \quad \sigma_{p_1} = \sqrt{\frac{(252)(.748)}{115}} = .0405, \text{ or } 4.05 \text{ per cent.}$$

$$\text{For Room 2:} \quad \sigma_{p_2} = \sqrt{\frac{(.080)(.920)}{112}} = .0256, \text{ or } 2.56 \text{ per cent.}$$

The standard error of the difference is

$$\sigma_{p_1 - p_2} = \sqrt{.0405^2 + .0256^2} = .0479, \text{ or } 4.79 \text{ per cent.}$$

Failures in Room 1 exceeded those in Room 2 by $25.2 - 8.0 = 17.2$ per cent, and

$$\frac{x}{\sigma} = \frac{p_1 - p_2}{\sigma_{p_1 - p_2}} = \frac{17.2}{4.79} = 3.59.$$

From the table of areas of the normal curve (Appendix E), it appears that, if the true difference between p_1 and p_2 is zero, a difference of 17.2 or more in favor of p_1 might occur about 1.5 times in 10,000. There appear to have been significantly fewer failures in Room 2, the difference presumably being attributable to better and more adaptable lighting.

When $N_1 \neq N_2$, instead of using two estimates of the p value for the population, we may combine the information available from the samples and make one estimate. This is a reasonable procedure since we set up the hypothesis that both samples are from the same universe. Denoting the combined estimate by p_{1+2},

$$p_{1+2} = \frac{N_1 p_1 + N_2 p_2}{N_1 + N_2},$$

and

$$\sigma'_{p_1 - p_2} = \sqrt{\frac{p_{1+2} \; q_{1+2}}{N_1} + \frac{p_{1+2} \; q_{1+2}}{N_2}},$$

---

[4] The data are from F. C. Albert, "Scholarship Improved by Light," *Transactions of the Illuminating Engineers Society.* December 1933, pp 866–872.

which is equivalent to

$$\sigma'_{p_1 - p_2} = \sqrt{p_{1+2}\, q_{1+2} \frac{N_1 + N_2}{N_1 N_2}}.$$

The procedure, from this point on, is the same as before.

Reliability of Measures of Dispersion

Reliability of a sample σ. We may test the significance of a sample σ in a manner similar to that previously described for $\overline{X}$. The standard error of the standard deviation σ_σ is given by

$$\sigma_\sigma = \frac{\sigma_P}{\sqrt{2N}}.$$

If kurtosis is present,

$$\sigma_\sigma = \frac{\sigma_P}{\sqrt{2N}} \sqrt{1 + \frac{\beta_2 - 3}{2}},$$

where β_2 is the kurtosis in the population

The standard deviation of the weights of 746 United States soldiers of French extraction was shown in Table 68 to be 16 03 pounds. At the time of demobilization, weight measurements were made not only of the United States soldiers of French extraction, but of over 80,000 soldiers of all types. For the entire group the value of σ_P was 17.06 pounds. Does the observed σ for the French differ significantly from this value? The value of $\sigma_\sigma = \dfrac{17.06}{\sqrt{2(746)}} = .4417$ pounds. The σ for the French troops was 1.027 pounds less than that for all troops. Comparing this difference with σ_σ gives

$$\frac{x}{\sigma} = \frac{\sigma - \sigma_P}{\sigma_\sigma} = \frac{1.027}{.4417} = 2.33.$$

Since N is large, we refer to the table of areas of the normal curve (Appendix E) and conclude that such a difference would rarely occur through chance arising from sampling.

If the value of σ_P is not known, we must substitute $\bar{\sigma}$ as computed from the sample and determine

$$\sigma_\sigma = \frac{\bar{\sigma}}{\sqrt{2N}}.$$

It will be recalled that $\sigma_{\overline{X}} = \dfrac{\bar{\sigma}}{\sqrt{N}}$, therefore

$$\sigma_\sigma = \frac{1}{\sqrt{2}} \sigma_{\overline{X}} = .7071068\, \sigma_{\overline{X}}.$$

Reliability of σ when N is small. Values of σ computed from small samples are not distributed normally or symmetrically. The distribution of sample values of σ^2 may be put in the form

$$\chi^2 = \frac{N\sigma^2}{\sigma_P^2}.$$

where the distribution of the population is assumed to be normal. Using this expression in conjunction with a table of χ^2 and using $n = N - 1$, we may ascertain the sampling variation of σ^2 or σ, provided we know the value of σ_P. Suppose that we have a sample of 10 items drawn from a population having $\sigma_P = 8$ pounds. What are the limits within which 98 per cent of the sample σ's (from samples of $N = 10$) would be expected to fall? It is apparent from the expression that a high value of σ^2 is associated with a high value of χ^2, and that a low value of σ^2 is associated with a low value of χ^2. We therefore determine the value of χ^2 at the .01 point and at the .99 point (since these limits include the central 98 per cent of the values of χ^2) and solve the expression above. Referring to the χ^2 table, for $n = 9$ and $P = .01$, we find $\chi^2 = 21.666$ and therefore

$$21.666 = \frac{10\sigma^2}{8^2}$$
$$10\sigma^2 = 1386.624$$
$$\sigma^2 = 138.6624$$
$$\sigma = 11.78 \text{ pounds.}$$

Referring to the χ^2 table, for $n = 9$ and $P = .99$, we find $\chi^2 = 2.088$ and

$$2.088 = \frac{10\sigma^2}{8^2}$$
$$10\sigma^2 = 133.632$$
$$\sigma^2 = 13.3632$$
$$\sigma = 3.66 \text{ pounds.}$$

From the above we conclude that sample σ's from samples of $N = 10$ would fall within the limits of 3.66 pounds and 11.78 pounds in 98 out of 100 instances.

The foregoing is useful if we know the value of σ_P. This will very rarely be true unless we have set up a control population (as is sometimes done in manufacturing) and desire to ascertain if samples selected from time to time correspond closely with this control group.

Ordinarily we know only the value of σ and N. When this is so, we may revert to the idea of fiducial probability and, using the same expression, ascertain the limits within which σ_P may confidently be expected to fall.

Let us determine the 90 per cent fiducial or confidence limits of σ_P for the hard-drawn copper wire previously referred to (Table 70). The value of σ was 8.26 pounds, and N was 10. Using the expression $\chi^2 = \dfrac{N\sigma^2}{\sigma_P^2}$, we proceed somewhat as before. At the .05 point, when $n = 9$, the value of χ^2 is 16.919 and

$$16.919 = \frac{10(8.26)^2}{\sigma_P^2}$$
$$16.919\sigma_P^2 = 682.276$$
$$\sigma_P^2 = 40.3260$$
$$\sigma_P = 6.35 \text{ pounds.}$$

At the .95 point, $\chi^2 = 3.325$ when $n = 9$ and

$$3.325 = \frac{10(8\ 26)^2}{\sigma_P^2}$$
$$3.325\sigma_P^2 = 682.276$$
$$\sigma_P^2 = 205.1958$$
$$\sigma_P = 14\ 32 \text{ pounds.}$$

There is a 5 per cent fiducial chance that σ_P is less than 6.35 pounds, and a 5 per cent fiducial chance that σ_P is greater than 14.32 pounds. The fiducial probability is 90 per cent that σ_P falls between 6.35 and 14.32 pounds.

Now let us ascertain the 98 per cent fiducial limits of σ_P for the copper wire. At the .01 point, when $n = 9$, the value of χ^2 is 21.666 and

$$21.666 = \frac{10(8.26)^2}{\sigma_P^2}$$
$$21.666\sigma_P^2 = 682.276$$
$$\sigma_P^2 = 31.4906$$
$$\sigma_P = 5.61 \text{ pounds.}$$

At the .99 point, $\chi^2 = 2.088$ when $n = 9$ and

$$2.088 = \frac{10(8.26)^2}{\sigma_P^2}$$
$$2.088\sigma_P^2 = 682.276$$
$$\sigma_P^2 = 326.7605$$
$$\sigma_P = 18.075 \text{ pounds.}$$

There is a 1 per cent fiducial chance that σ_P is less than 5.61 pounds, and a 1 per cent fiducial chance that σ_P is greater than 18.075 pounds. The fiducial probability is 98 per cent that σ_P lies between 5.61 and 18.075 pounds. If we need to reduce the fiducial limits of σ_P, we must study a' larger sample.

Considering the expression $\chi^2 = \dfrac{N\sigma^2}{\sigma_P^2}$, it will be apparent that, for given values of χ^2 and N, the ratio $\dfrac{\sigma^2}{\sigma_P^2}\left(\text{or } \dfrac{\sigma_P^2}{\sigma^2}\right)$ will always be a constant, as will also the ratio $\dfrac{\sigma}{\sigma_P}\left(\text{or } \dfrac{\sigma_P}{\sigma}\right)$. For any given level of fiducial probability and any given sample size, it is possible to ascertain this ratio. Since we are interested in inferring the value of σ_P from a known value of σ, we shall consider the ratio $\dfrac{\sigma_P}{\sigma}$. Let us call this ratio b $\left(\text{that is, } b = \dfrac{\sigma_P}{\sigma}\right)$, and write $\sigma_P = b\sigma$.

Suppose we wish to determine the values of b for the .05 and .95 level when $N = 10$ $(n = 9)$. Referring back to the illustration of the hard-drawn copper wire, we found that there was a 90 per cent fiducial probability that σ_P fell between 6.35 pounds and 14.32 pounds, while σ was 8.26 pounds. Using b_1 for the lower fiducial value, we have

$$b_1 = \frac{6.35}{8.26} = .769.$$

Using b_2 for the upper fiducial value gives

$$b_2 = \frac{14.32}{8.26} = 1.734.$$

For samples of $N = 10$, there is a 90 per cent fiducial probability that σ_F falls between $b_1\sigma$ and $b_2\sigma$. In similar fashion, the values of b_1 and b_2 for samples of various sizes could be determined. A number of these values are given in Table 72 and enable us quickly to ascertain the 90 per cent fiducial limits of σ_P.

The values of b_3 and b_4 for the 98 per cent fiducial limits of σ_P when $N = 10$ may also be ascertained. We found a 98 per cent fiducial probability that σ_P for the hard-drawn copper wire fell between 5.61 and 18.075 pounds. Then

$$b_3 = \frac{5.61}{8.26} = .679.$$

$$b_4 = \frac{18.075}{8.26} = 2.188.$$

For samples of $N = 10$, there is a 98 per cent fiducial probability that σ_P falls between $b_3\sigma$ and $b_4\sigma$. Similarly, values of b_3 and b_4 may be computed for samples of various sizes, and a number of these are given in Table 72 for the 98 per cent fiducial limits of σ_P.

TABLE 72

VALUES OF b_1, b_2, b_3, AND b_4 FOR DETERMINING FIDUCIAL OR CONFIDENCE LIMITS OF σ_P FOR SAMPLES OF $N = 5$ TO $N = 30$

N	90 per cent fiducial limits		98 per cent fiducial limits	
	b_1	b_2	b_3	b_4
5	726	2 652	.614	4.103
6	.736	2 289	631	3 291
7	.746	2 069	645	2 833
8	.754	1 921	.658	2 541
9	.762	1 815	669	2 338
10	.769	1 734	679	2 188
11	.775	1.671	688	2 074
12	781	1 620	697	1 983
13	.786	1.577	704	1.908
14	791	1 541	.711	1 846
15	.796	1.511	.717	1.794
16	800	1.484	.723	1 749
17	.804	1 461	729	1.710
18	.808	1 441	.734	1.676
19	.811	1 422	739	1.646
20	.815	1.406	.743	1.619
21	.818	1.391	.748	1 594
22	.821	1.378	752	1.572
23	.823	1 365	.756	1 553
24	826	1 354	.759	1.534
25	829	1.344	.763	1.518
26	.831	1.334	766	1.502
27	.833	1 325	769	1 488
28	.835	1 317	.772	1.474
29	.838	1 309	.775	1.462
30	.840	1 302	778	1.451

Significance of difference between two standard deviations when N's are large and when $N_1 = N_2$. For the groups of left-handed and right-handed students discussed earlier, it was found that no significant difference existed between the two mean I.Q.'s. For left-handed students, σ_1 was 15.1; while for right-handed students, σ_2 was 15.4. Is the difference between these measures significant? It will be remembered that $\bar{\sigma}_1 = 15.2$, $\bar{\sigma}_2 = 15.5$, and $N_i = N_2 = 68$.

$$\sigma_{\sigma_1} = \frac{\bar{\sigma}_1}{\sqrt{2N}} = \frac{15.2}{\sqrt{136}} = 1.30.$$

$$\sigma_{\sigma_2} = \frac{\bar{\sigma}_2}{\sqrt{2N}} = \frac{15.5}{\sqrt{136}} = 1.33.$$

$$\sigma_{\sigma_1 - \sigma_2} = \sqrt{\sigma_{\sigma_1}^2 + \sigma_{\sigma_2}^2} = \sqrt{1.30^2 + 1.33^2} = 1.86.$$

The ratio of the observed difference to the standard error of the difference is

$$\frac{x}{\sigma} = \frac{\sigma_1 - \sigma_2}{\sigma_{\sigma_1 - \sigma_2}} = \frac{15.1 - 15.4}{1.86} = \frac{.3}{1.86} = .16.$$

Referring to the table of areas of the normal curve, we find that in 872 instances out of 1,000 we might expect to get a difference of $\pm.3$ or more between the σ's through the variations of sampling, or that σ_2 would exceed σ_1 by .3 in 436 cases out of 1,000. We conclude that there is not a significant difference between the two σ's.

It will be apparent from the expressions for $\sigma_{\bar{x}}$ and σ_σ that

$$\sigma_{\sigma_1 - \sigma_2} = \frac{1}{\sqrt{2}} \sigma_{\bar{x}_1 - \bar{x}_2} = .7071068\, \sigma_{\bar{x}_1 - \bar{x}_2}.$$

Referring to page 319, it was found that $\sigma_{\bar{x}_1 - \bar{x}_2}$ for the I.Q.'s of left-handed and right-handed students was 2.63, and $.7071068 \times .263 = 1.86$, which is the same as computed above.[5]

Significance of differences between two σ's when N's are small and/or $N_1 \neq N_2$. When N_1 and N_2 are both large, or when N_1 and N_2 are moderate values and $N_1 = N_2$ or nearly so, we may proceed as above. It was previously noted that the sampling distribution of σ is not normal when N is small. To test the significance of differences between standard deviations, R. A. Fisher has suggested a transformation which is particularly useful for small samples and which refers to $\bar{\sigma}_1$ and $\bar{\sigma}_2$ instead of σ_1 and σ_2. This transformation is:

$$z = (\log_e \bar{\sigma}_1 - \log_e \bar{\sigma}_2) = \log_e \frac{\bar{\sigma}_1}{\bar{\sigma}_2},$$

or

$$z = \tfrac{1}{2} (\log_e \bar{\sigma}_1^2 - \log_e \bar{\sigma}_2^2) = \tfrac{1}{2} \log_e \frac{\bar{\sigma}_1^2}{\bar{\sigma}_2^2}.$$

---

[5] The standard error of the coefficient of variation $(V = \frac{\sigma}{\bar{X}})$ is

$$\sigma_V = \frac{V_P}{\sqrt{2N}} \sqrt{1 + 2V_P^2},$$

where V_P (expressed as a decimal) refers to the coefficient of variation of the population. When $N_1 = N_2$, $\sigma_{V_1 - V_2} = \sqrt{\sigma_{V_1}^2 + \sigma_{V_2}^2}.$

Since adequate tables of natural logarithms are not available, we may compute the value of z by making use of the expression

$$\log_e X = \log_e 10 \cdot \log_{10} X = 2.302585 \log_{10} X.$$

Therefore

$$z = 2.302585 \log_{10} \frac{\bar{\sigma}_1}{\bar{\sigma}_2},$$

or

$$z = 1.15129 \log_{10} \frac{\bar{\sigma}_1^2}{\bar{\sigma}_2^2}.$$

Use of z enables us to test the reliability of the difference between $\bar{\sigma}_1$ and $\bar{\sigma}_2$ or between $\bar{\sigma}_1^2$ and $\bar{\sigma}_2^2$ It must be obvious that the two expressions are exactly the same; if there is a significant difference between $\bar{\sigma}_1$ and $\bar{\sigma}_2$, there is also a significant difference between $\bar{\sigma}_1^2$ and $\bar{\sigma}_2^2$.

The value of z varies between plus and minus infinity, being negative when

$$\frac{\bar{\sigma}_1^2}{\bar{\sigma}_2^2} < 1,$$

and positive when

$$\frac{\bar{\sigma}_1^2}{\bar{\sigma}_2^2} > 1.$$

The distribution of z is approximately normal when N_1 and N_2 are both large, or when N_1 and N_2 are moderate and equal or nearly equal.[6] Unless

---

[6] See L H C. Tippett, *The Methods of Statistics*, pp. 117–120, Williams and Norgate, London, 1937 (2nd Edition).

We have studied the reliability of σ^2 by making use of the distribution of χ^2. Now we shall study the significance of differences between variances by transforming the computed values into z values   Sometimes we must work with a distribution the exact shape of which is not known   For any series of values, no matter how distributed, it may be shown by *Tchebycheff's inequality* that the proportion of values lying beyond a given symmetrical range of the mean $\pm \frac{x}{\sigma}$ is less than $\dfrac{1}{\left(\frac{x}{\sigma}\right)^2}$. That is, $P < \dfrac{1}{\left(\frac{x}{\sigma}\right)^2}$. This is an extremely conservative test of reliability   As is apparent from the expression $\dfrac{1}{\left(\frac{x}{\sigma}\right)^2}$, the .05 level of significance ($P < .05$) is at $4\,47\sigma$; the 02 level ($P < .02$) is at 7.07σ; the .01 level ($P < .01$) is at 10σ; and the .001 level ($P < .001$) is at 31.62σ

If a distribution is unimodal, and if the mode is within σ of the mean [that is, if

$N_1 = N_2$, the distribution of z is skewed and, for simplicity of procedure. we generally work only with positive values of z, which is accomplished by considering the series with the larger standard deviation (or variance) as the first series with subscript 1. This makes it necessary to consider only the positive half of the skewed distribution of z. The significance of z depends upon the value computed for z and also upon $n_1(= N_1 - 1)$ and $n_2(= N_2 - 1)$. To present a reasonably complete table of the distribution of z would require many pages. However, Fisher, has prepared tables for $P = .05$, $P = .01$, and $P = .001$ points for selected values of n_1 and n_2. These are shown as Appendix G1.

As an illustration let us compare the variance of length of the two groups of lower first molars previously mentioned. Notice that we are comparing $\bar{\sigma}_1^2$ and $\bar{\sigma}_2^2$ rather than σ_1^2 and σ_2^2.

For the molars excavated at the first site:

$$\Sigma x_1^2 = 8.2944.$$
$$n_1 = N_1 - 1 = 16 - 1 = 15.$$
$$\bar{\sigma}_1^2 = \frac{8.2944}{15} = .553.$$

For the molars excavated at the second site:

$$\Sigma x_2^2 = 3.4596.$$
$$n_2 = N_2 - 1 = 9 - 1 = 8.$$
$$\bar{\sigma}_2^2 = \frac{3.4596}{8} = .432.$$

The computation of z is based upon the foregoing.

$$z = 1.15129 \log_{10} .553 - 1.15129 \log_{10} .432$$
$$= 1.15129(9.742725 - 10) - 1.15129(9.635484 - 10)$$
$$= .1235.$$

---

skewness as measured by (mean $-$ mode) $\div \sigma \leq \pm 1$] we may apply the *Camp-Meidell*

inequality, which states that *less than* $\dfrac{1}{2.25\left(\dfrac{x}{\sigma}\right)^2}$ of the items $\left[P < \dfrac{1}{2.25\left(\dfrac{x}{\sigma}\right)^2}\right]$ lie beyond

a given symmetrical range of the mean $\pm\dfrac{x}{\sigma}$. Upon this basis, the 05 level ($P < 05$) is at 2.98σ; the .02 level ($P < .02$) is at 4.71σ; the .01 level ($P < .01$) is at $6\ 67\sigma$; and the .001 level ($P < .001$) is at $21\ 08\sigma$. See W. A. Shewhart, *Economic Control of Manufactured Product*, pp 175–176, D. Van Nostrand Co., New York, 1931; and B H. Camp, *The Mathematical Part of Elementary Statistics*, pp 256–257, D. C. Heath. Boston, 1931.

Alternatively,

$$z = 1.15129 \log_{10} \frac{.553}{.432}$$
$$= 1.15129 \log_{10} 1.2801$$
$$= .1235.$$

Referring to the z table (Appendix G1), for $n_1 = 15$ and $n_2 = 8$, we find that a value of z of about .58 falls at the .05 point, while a value of z of about .85 falls at the .01 point. Consequently the chances of obtaining a value of $z = .1235$ are appreciably greater than .05, and it appears that there is not a significant difference between the two variances. It must be apparent from the expression

$$z = \tfrac{1}{2} \log_e \frac{\bar{\sigma}_1^2}{\bar{\sigma}_2^2}$$

that the value of z depends upon the ratio of $\bar{\sigma}_1^2$ to $\bar{\sigma}_2^2$ and not upon the absolute value of either variance. For example, if $\bar{\sigma}_1^2 = 4$ and $\bar{\sigma}_2^2 = 2$, then $z = \tfrac{1}{2} \log_e 2 = .69315$. Similarly, if $\bar{\sigma}_1^2 = 21.6$ while $\bar{\sigma}_2^2 = 10.8$, the value of $z = \tfrac{1}{2} \log_e 2 = .69315$, as before. For this reason it is possible to recast the table of z in Appendix G1, and state it in terms of $F = \dfrac{\bar{\sigma}_1^2}{\bar{\sigma}_2^2}$ as shown in Appendix G2. The larger variance should always be in the numerator when we use this table. The F table is a more convenient table to use since it eliminates the necessity of looking up logarithms. If we wish to use this table for testing the significance of the difference between $\bar{\sigma}_1$ and $\bar{\sigma}_2$, we determine $F = \dfrac{\bar{\sigma}_1^2}{\bar{\sigma}_2^2}$, or we may compute $\dfrac{\bar{\sigma}_1}{\bar{\sigma}_2}$ and square the resulting figure before entering the table.

The z test may, of course, be used when the N's are large. Let us apply it to the data of I.Q.'s of right- and left-handed students previously discussed. For 68 left-handed students, we found $\bar{\sigma}_1 = 15.2$; for 68 right-handed students, $\bar{\sigma}_2 = 15.5$. We can now proceed as before when computing z, and interpolate in Fisher's table for $n_1 = 67$ and $n_2 = 67$. But the sampling distribution of z is approximately normal when N_1 and N_2 are both large, or when N_1 and N_2 are moderate and equal or nearly so. It is therefore substantially accurate in the present instance to find the ratio of z to σ_z, and interpret by reference to the normal curve. Since the distribution of z is approximately normal and not skewed as in the preceding illustration, it is not necessary that z be positive.

$$z = 2.30259 \log_{10} 15.2 - 2.30259 \log_{10} 15.5$$
$$= 2.30259(1.181844) - 2.30259(1.190332)$$
$$= .01954.$$

However, if we should compute z from the expression

$$z = 2.30259 \log_{10} \frac{\bar{\sigma}_1}{\bar{\sigma}_2},$$

it would be necessary to consider the larger $\bar{\sigma}$ as $\bar{\sigma}_1$.

The value of σ_z is given by:

$$\sigma_z = \sqrt{\frac{1}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$= \sqrt{\frac{1}{2}\left(\frac{1}{67} + \frac{1}{67}\right)} = \sqrt{\frac{1}{67}}$$

$$= \sqrt{.01493} = .1222.$$

The ratio

$$\frac{z}{\sigma_z} = \frac{.01954}{.1222} = .16$$

is the same as that obtained when comparing σ_1 and σ_2 on page 344, and indicates no significant difference between $\bar{\sigma}_1$ and $\bar{\sigma}_2$.

An Application of Reliability Measures

Control of quality of manufactured product may involve the adherence to a predetermined standard in order to maintain at all times a high degree of uniformity of product. Suppose a manufacturer wishes to control his production of $\frac{3}{8}$-inch bolts so that, among other characteristics, the breaking strength will not be less than 6,000 pounds. He cannot, of course, maintain an entirely uniform product. Causes that affect the uniformity of tensile strength are variations in the carbon content of the steel, in impurities such as sulphur and phosphorous, in homogeneity of structure, in conditions of heat treatment, in the diameter of the rod stock, and in the manufacturing process. If each of the causes of variation continues to have the same probability of contributing a given effect, then the breaking strength may be said to be controlled in a technical sense. Thus, by keeping the conditions of manufacture under control, an acceptable and relatively uniform product may be assured.

If the cost of testing each bolt separately is prohibitive, or if the test is destructive, it will be necessary to resort to sampling in order to ascertain whether quality and uniformity are being maintained.[7] In order to estab-

---

[7] Except that in certain instances an associated characteristic may be tested, as for example, hardness may be used as an indicator of tensile strength. See F. E. Croxton and D. J. Cowden, *Practical Business Statistics*, pp. 405–416, Prentice Hall, Inc., New York, 1934.

lish statistical control, it is obviously necessary to set a standard level for the average tensile strength of bolts. The first step, then, is to determine the arithmetic mean and the standard deviation of tensile strength of such bolts produced under standard conditions, and these measures are taken as the mean and the standard deviation of the entire population. If not more than about 1 bolt out of 1,000 is to be less than 6,000 pounds in tensile strength, it is necessary that the manufacturing process be established so that the average tensile strength will be larger than 6,000 by an amount equal to at least three standard deviations, since the tail of a normal curve beyond -3σ contains about $\frac{13}{1000}$ of the area.

Assuming that $\sigma_P = 320$ pounds, $3\sigma_P$. $=$ 960 pounds
Lowest permissible value for tensile strength $=$ 6,000 pounds

Required $\overline{X}_P$ . .. . $=$ 6,960 pounds.

It is now possible to estimate the limits within which the means of any given proportion of the samples of size N should fall if drawn from the standard universe of bolts. These limits are, of course, $\overline{X}_P \pm \sigma_{\overline{x}}$ for the means of 68.27 per cent of the samples, and $\overline{X} \pm 3\sigma_{\overline{x}}$ for 99.73 per cent of the sample means. If each sample is to be of 4 bolts,

$$\sigma_{\overline{x}} = \frac{320}{\sqrt{4}} = 160 \text{ pounds.}$$

Since we have σ_P instead of $\bar{\sigma}$, we make use of the areas under the normal curve rather than the t curve.

Therefore, 99.73 per cent of the means of samples of 4 bolts each should vary between $6,960 \pm 3(160)$, and only about 1 out of every 1,000 will be less than $6,960 - 3(160) = 6,480$ if they are drawn from this universe. If a sample mean is less than 6,480, as in Chart 136, this result indicates that it was probably not drawn from the standard universe, and that therefore lack of control exists. The cause of the difficulty should be traced. Had the sample been of 16 rather than of 4, the allowable limit for a sample mean would have been only half as far below the standard mean, since the reliability of a sample increases as the square root of the number of items included. Engineers have adopted the 3σ limits as an indication of lack of control, not so much because of the statistical probability value of 99.73 per cent as because these limits have proved to be satisfactory and economic in practice.

Samples indicate probability only, not certainty. It is quite possible that lack of manufacturing control might exist without being brought to light by this procedure. It is even possible that a sample mean might
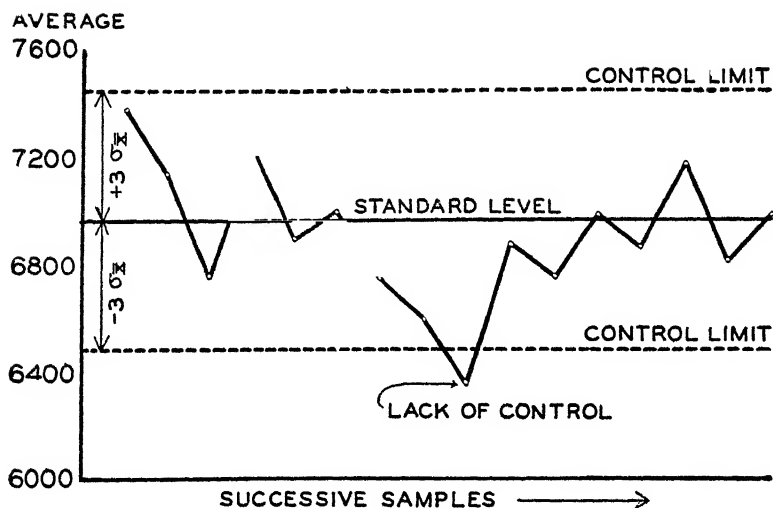
Chart 136. Arithmetic Means of the Tensile Strength of Successive Test Samples Each of Four ⅜-Inch Bolts. (Based on a chart from H. F. Dodge, "Statistical Control in Sampling Inspection," *American Machinist*, October 26 and November 9, 1932.)
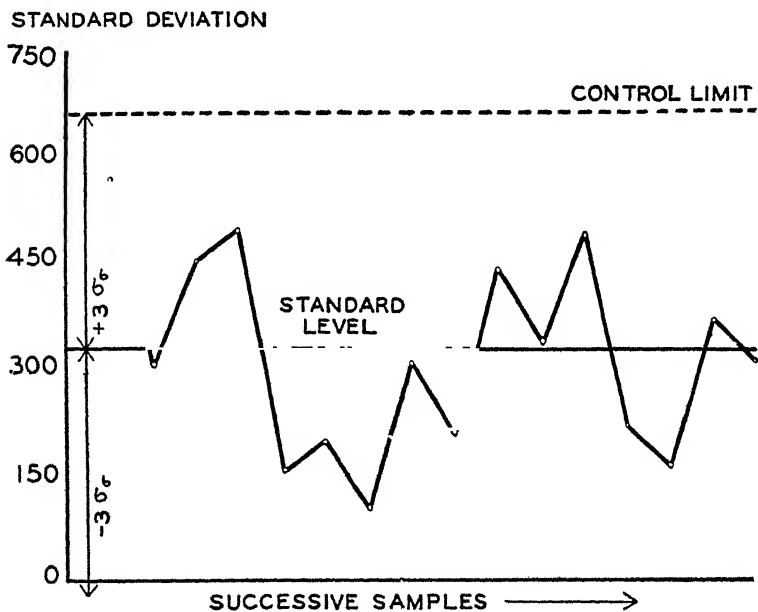
Chart 137. Standard Deviations of the Tensile Strength of Successive Test Samples Each of Four ⅜-Inch Bolts. (Based on chart from H. F. Dodge, "Statistical Control in Sampling Inspection," *American Machinist*, October 26 and November 9, 1932.)

be well above the control limit and yet that individual bolts in that sample might have a tensile strength of less than 6,000 pounds; this could easily be true if the variation within a sample was large. To further assure satisfactory quality as well as a uniform product, it is desirable also to set up in analogous fashion control limits for the standard deviation of samples and to record the results of successive samples as in Chart 137.[8]

Analysis of Variance

The analysis of variance provides a basis for comparing not only two, but any number of series simultaneously.

In Table 73 data are shown of the length of English cuckoo eggs which were deposited in the nests of three different species of birds. The cuckoo makes a practice of permitting other birds to hatch its eggs and rear its offspring. We are interested in knowing if the lengths of cuckoo eggs laid in the nests of the tree-pipit, the pied wagtail, and the wren show significant differences. We could, of course, compute the three means and compare the first mean with the second, the first with the third, and the second with the third. This, however, does not measure the variability of the three groups as a whole.

Following the concept mentioned in Chapter IV when constructing line diagrams, we shall regard the type of nest as the X variable, or independent variable, and the length of eggs as the Y variable, or dependent variable. There were fifteen eggs measured from each of the three types of nests, a total of 45 measurements as shown in the table. The *total variation* present in the entire series of measurements is measured by the sum of the squared deviation of each measurement from the mean of all 45 measurements. Thus the total variation is

$$\Sigma \, y^2 = \Sigma \, (Y - \overline{Y})^2.$$

The *variation within the groups* is measured by considering the deviation of each measurement from the mean of its group; these deviations are then squared and summed. Letting $\overline{Y}_1$ refer to the mean of the first column, and $\overline{Y}_2$ and $\overline{Y}_3$ the means of the succeeding columns, we have:

---

[8] This application of probability theory is summarized from a paper presented by H. F. Dodge of the Bell Telephone Laboratories on "Statistical Control in Sampling Inspection," at the annual meeting of the American Society for Testing Materials, Atlantic City, June 20, 1932, and published in *American Machinist*, October 26 and November 9, 1932   See also W. A. Shewhart, *Economic Control of Quality of Manufactured Product*, D. Van Nostrand Co., New York, 1931; and *1933 A. S. T. M. Manual on Presentation of Data*, 1933, with Supplements A (Presenting Plus and Minus Limits of Uncertainty of an Observed Average) and B ("Control Chart" Method of Analysis and Presentation of Data), 1935. American Society for Testing Materials, Philadelphia.

TABLE 73

LENGTH OF CUCKOO'S EGGS DEPOSITED IN THE NESTS OF THREE SPECIES OF BIRDS

Order of measure-ment	Tree-pipit		Pied Wagtail		Wren	
	Length in millimeters Y_1	Y_1^2	Length in millimeters Y_2	Y_2^2	Length in millimeters Y_3	Y_3^2
1	22 7	515 29	23 0	529 00	19.8	392 04
2	23 3	542 89	23 4	547 56	22 1	488 41
3	24 0	576 00	24 0	576 00	21 5	462.25
4	23 6	556 96	23 3	542 89	20 9	436 81
5	22 1	488 41	23 1	533 61	22 0	484 00
6	21 8	475.24	22 4	501 76	21 0	441 00
7	21 1	445 21	21 8	475 24	22 3	497 29
8	23 4	547 56	21 8	475 24	21 0	441 00
9	23 8	566 44	24 9	620 01	20 3	412 09
10	23.3	542 89	24.0	576.00	20.9	436 81
11	24 0	576 00	22 1	488 41	22 0	484 00
12	23 5	552 25	21 0	441 00	20 0	400 00
13	23 2	538 24	22 6	510.76	20 8	432 64
14	24 0	576 00	21.9	479 61	21 2	449 44
15	22 4	501 76	24.0	576 00	21 0	441 00
Total	346.2	8,001.14	343.3	7,873 09	316 8	6,698.78
	ΣY_1	ΣY_1^2	ΣY_2	ΣY_2^2	ΣY_3	ΣY_3^2

Source Oswald H Latter, "The Egg of Cuculus Canorus," *Biometrika*, Vol I, pp 164–176.

$$\Sigma Y = 346\ 2 + 343.3 + 316\ 8 = 1,006\ 3$$
$$(\Sigma Y)^2 = (1,006\ 3)^2 = 1,012,639.69$$
$$\Sigma Y^2 = 8,001\ 14 + 7,873\ 09 + 6,698.78 = 22,573\ 01$$
$$\sum_1^m \left(\sum_1^{N_K} Y \right)^2 = (\Sigma Y_1)^2 + (\Sigma Y_2)^2 + (\Sigma Y_3)^2$$
$$= (346\ 2)^2 + (343\ 3)^2 + (316\ 8)^2 = 338,071.57.$$

Degrees of freedom:

Between nests (columns). ... 2 $N_K = 15$

Within nests (columns).. . . 42 $N = 45$

— $m = 3$

Total...... 44

For the first column: $\displaystyle\sum_{1}^{N_K} (Y - \overline{Y}_1)^2$

For the second column: $\displaystyle\sum_{1}^{N_K} (Y - \overline{Y}_2)^2$

For the third column: $\displaystyle\sum_{1}^{N_K} (Y - \overline{Y}_3)^2$

For all columns: $\displaystyle\sum_{1}^{m}\left[\sum_{1}^{N_K} (Y - \overline{Y}_K)^2\right]$

where N_K represents the number of items in a column;

$\overline{Y}_K$ represents the mean of a column;

$\displaystyle\sum_{1}^{N_K}$ indicates a summation for a column;

$\displaystyle\sum_{1}^{m}$ indicates a summation of the m columns;

$\sum$ indicates, as before, a summation for the entire series.

The *variation between the groups* is measured by referring to the deviation of each group mean from the grand mean (the mean of all the data); these deviations are squared, multiplied by the number of items in the group, and summed. Thus:

For the first column: $N_1(\overline{Y}_1 - \overline{Y})^2$

For the second column: $N_2(\overline{Y}_2 - \overline{Y})^2$

For the third column: $N_3(\overline{Y}_3 - \overline{Y})^2$

For all columns: $\displaystyle\sum_{1}^{m}\left[N_K(\overline{Y}_K - \overline{Y})^2\right]$

It is shown in Appendix B[9] that

$$\sum (Y - \overline{Y})^2 = \sum_{1}^{m}\left[\sum_{1}^{N_K} (Y - \overline{Y}_K)^2\right] + \sum_{1}^{m}\left[N_K (\overline{Y}_K - \overline{Y})^2\right],$$

or, in other words, that total variation = variation within columns + variation between columns.

If we divide the variation within columns by the degrees of freedom present ($14 + 14 + 14 = 42$, in this case), we obtain a measure of the *variance*[10] *within columns*. If we divide the variation between columns

---

[9] Appendix B, section XIII-2.

[10] In computing variance in this discussion, we always use degrees of freedom. Thus all variances are estimates of population variance rather than statements of the sample variance. In the later chapters on correlation, we shall make use of measures of variance based upon N rather than n.

by the degrees of freedom (2 in this instance), we obtain a measure of the *variance between columns*. Now the variance within columns (within nests) is clearly a chance variation. If the variance between columns (between types of nests) exceeds the former significantly, then there is a significant difference in length of eggs in these three types of nests. Or, it may be said that the total variance $\left(\dfrac{\Sigma y^2}{N-1}\right)$ in length of eggs is partly explained by the types of nests in which they are found, while the variance within nests is a chance variation because it has not been explained, nor, in fact, did the hypothesis even attempt to explain it.

The variation within columns $\sum\limits_{1}^{m}\left[\sum\limits_{1}^{N_K}(Y-\overline{Y}_K)^2\right]$ could be computed by determining the mean of each column, taking the deviation of each of the 45 measurements from the appropriate column mean, squaring, and adding. In computing σ, we developed a short method which made it unnecessary to work with deviations. A similar device may be employed here, and in Appendix B (section XIII-2) it is shown that

$$\sum_{1}^{m}\left[\sum_{1}^{N_K}(Y-\overline{Y}_K)^2\right]=\Sigma Y^2-\frac{\sum\limits_{1}^{m}\left(\sum\limits_{1}^{N_K}Y\right)^2}{N_K}$$

if there is the same number of items in the various columns. Referring to the computed values shown below Table 73, we find

$$\Sigma Y^2-\frac{\sum\limits_{1}^{m}\left(\sum\limits_{1}^{N_K}Y\right)^2}{N_K}=22{,}573.01-\frac{338{,}071.57}{15}$$

$$=22{,}573.01-22{,}538.10$$

$$=34.91.$$

This figure is entered in Table 74 as the variation or sum of the squared deviations within columns. The mean variance or merely *variance within columns*, as it is usually termed, is obtained by dividing this figure by the degrees of freedom. Since there are 15 items in each column and since the squared deviations were taken in reference to the mean of each column, there are 14 degrees of freedom in each column or $3 \times 14 = 42$ degrees of freedom within columns. The variance within columns is $\dfrac{34.91}{42} = .8312$, which is also shown in Table 74.

The variation between columns $\sum\limits_{1}^{m}\left[N_K(\overline{Y}_K-\overline{Y})^2\right]$ may also be obtained

TABLE 74

ANALYSIS OF VARIANCE OF THE LENGTH OF CUCKOO'S EGGS DEPOSITED IN NESTS OF THREE SPECIES OF BIRDS

Source of variation	Variation or sum of squared deviations	Degrees of freedom	Variance
Within nests (i e , within columns of Table 73). . .	34 91	42	8312
Between nests (i e , between columns of Table 73)	35 00	2	17.50
Total	69 91	44	...

The total variation was computed from

$$\Sigma(Y - \overline{Y})^2 = \Sigma Y^2 - \frac{(\Sigma Y)^2}{N}$$

See Appendix B, section XIII–2   Total variation is shown in this table for checking purposes, since it is the total of the two figures above it   Likewise, total degrees of freedom $(N - 1)$ is the sum of the two figures preceding

$$\Sigma Y^2 - \frac{(\Sigma Y)^2}{N} = 22{,}573\ 01 - \frac{1{,}012{,}639.69}{45}$$

$$= 22{,}573.01 - 22{,}503\ 10$$

$$= 69\ 91.$$

Total variance, however, is not the sum of the other two variances.   It may be computed by dividing total variation by total degrees of freedom.   Thus

$$\frac{69\ 91}{44} = 1.589.$$

without computing means or deviations.   It is shown in Appendix B[11] that

$$\sum_{1}^{m}\left[N_K(\overline{Y}_K - \overline{Y})^2\right] = \frac{\sum_{1}^{m}\left(\sum_{1}^{N_K} Y\right)^2}{N_K} - \frac{(\Sigma Y)^2}{N}$$

if there is the same number of items in each column.   Again referring to Table 74, we have

$$\frac{\sum_{1}^{m}\left(\sum_{1}^{N_K} Y\right)^2}{N_K} - \frac{(\Sigma Y)^2}{N} = \frac{338{,}071.57}{15} - \frac{1{,}012{,}639.69}{45}$$

$$= 22{,}538.10 - 22{,}503.10$$

$$= 35.00.$$

---

[11] Appendix B, section XIII–2.

This figure is entered in Table 74 as the variation or sum of the squared deviations between columns. The *variance between columns* is now obtained by dividing this figure by the degrees of freedom. There are 2 degrees of freedom between columns since the 3 column means were considered in relation to the grand mean, thus losing 1 degree of freedom The variance between columns is $\frac{35\ 00}{2} = 17.50$, which is also entered in Table 74.

The variance within nests (columns) is .8312, while the variance between nests (columns) is 17.50. As previously mentioned, the variance within nests (columns) can logically be considered as due to chance causes. If the variance between nests does not exceed the variance within nests significantly, we may conclude that the variance between nests also is due to chance. The variance between nests is much greater than the variance within nests and hardly needs to be tested for significance   The test which is used is the same z test previously applied to two values of $\bar{\sigma}^2$ when $N_1 \neq N_2$. It will be remembered that the larger variance appears in the numerator in the expression for z, in order that the value of z will be positive.

$$z = 1.15129 \log_{10} \frac{17.50}{.8312} = 1.15129 \log_{10} 21.05$$

$$= 1.52345.$$

Referring to Appendix G1, mentioned before, and using degrees of freedom $n_1 = 2$ and $n_2 = 42$, we find that $z = 1.52345$ lies beyond the $\frac{1}{10}$ of one per cent point and the difference is clearly significant. The variation of egg length with type of nest is therefore almost certainly real.[12]   Computing $F = \frac{17.50}{.8312} = 21.05$ and referring to Appendix G2 for $n_1 = 2$ and $n_2 = 42$, the conclusion is, of course, exactly the same as that arrived at by use of the z test.

In Table 75, data are given of the strength of the lead in five pencils made by a company identified only as "Company D." Four tests were made of the lead of each pencil. We are interested in knowing if the variance between pencils is significantly greater than the variance within pencils, in order to find out whether or not there is uniformity of strength from pencil to pencil.

---

[12] Compare with Tippett (*The Methods of Statistics*, pp. 132–134, 2nd Edition), who comes to the same conclusion using 6 nest-types with N_K ranging from 14 to 45.

TABLE 75

Strength of Lead in Number 2 Pencils Manufactured by "Company D"

Pencil 1		Pencil 2		Pencil 3		Pencil 4		Pencil 5	
Strength in kilograms Y_1	Y_1^2	Strength in kilograms Y_2	Y_2^2	Strength in kilograms Y_3	Y_3^2	Strength in kilograms Y_4	Y_4^2	Strength in kilograms Y_5	Y_5^2
1.82	3 3124	1 70	2 8900	1 70	2 8900	1 82	3 3124	1 92	3 6864
1.56	2 4336	1.36	1 8496	1 68	2 8224	1 98	3 9204	1 86	3 4596
1.78	3.1684	1 54	2 3716	2 02	4 0804	1 82	3 3124	1 64	2 6896
1.74	3.0276	1 92	3 6864	1 92	3 6864	1 64	2 6896	1 75	3 0625
6 90	11 9420	6 52	10 7976	7 32	13 4792	7 26	13 2348	7 17	12 8981

$N_K = 4$
$N = 20$
$m = 5$

Source: From tests of pencils of various brands conducted in 1934 by the Eagle Pencil Co

Degrees of freedom:

Between pencils (columns). 4
Within pencils (columns). 15

Total...... 19

$\Sigma Y = 35\ 17$

$\Sigma Y^2 = 62\ 3517$

$(\Sigma Y)^2 = 1{,}236\ 9289.$

$\sum_{1}^{m} \left(\sum_{1}^{N_K} Y \right)^2 = 247\ 8193.$

357

Variation within pencils (columns) is

$$\Sigma Y^2 - \frac{\overset{m}{\underset{1}{\Sigma}}\left(\overset{N_K}{\underset{1}{\Sigma}} Y\right)^2}{N_K} = 62.3517 - \frac{247.8193}{4}$$

$$= 62.3517 - 61.954825$$
$$= .396875.$$

Variance within pencils (5×3 degrees of freedom) is

$$\frac{.396875}{15} = .02646.$$

Variation between pencils (columns) is

$$\frac{\overset{m}{\underset{1}{\Sigma}}\left(\overset{N_K}{\underset{1}{\Sigma}} Y\right)^2}{N_K} - \frac{(\Sigma Y)^2}{N} = \frac{247.8193}{4} - \frac{(35.17)^2}{20}$$

$$= 61.954825 - 61.846445$$
$$= .108380.$$

Variance between pencils (4 degrees of freedom) is

$$\frac{.108380}{4} = .027095.$$

These figures are summarized in Table 76 together with the figures for total variation and total degrees of freedom for purposes of checking. It

TABLE 76

ANALYSIS OF VARIANCE OF STRENGTH OF LEAD IN NUMBER 2 PENCILS MANUFACTURED BY "COMPANY D"

Source of variation	Variation or sum of squared deviations	Degrees of freedom	Variance
Within pencils.	.396875	15	026458
Between pencils	.108380	4	027095
Total	.505255	19	

Total variation is computed from

$$\Sigma Y^2 - \frac{(\Sigma Y)^2}{N} = 62.3517 - \frac{1236.9289}{20}$$

$$= 62.3517 - 61.846445$$
$$= .505255.$$

will be observed that the two variances are very nearly alike. Making the z test,

$$z = 1.15129 \log_{10} \frac{.027095}{.026458} = 1.15129 \log_{10} 1.024$$

$$= .01186.$$

n_1 (degrees of freedom between pencils) is 4, while n_2 (degrees of freedom within pencils) is $5 \times 3 = 15$. Consulting Appendix G1, it appears that when $n_1 = 4$ and $n_2 = 15$, a value of $z = .5585$ falls on the .05 level of significance. Since the z obtained above is much less than this, it cannot be said that there is a significant difference between these variances. Instead of using z, we may use $F = \frac{.027095}{.026458} = 1.024$, $n_1 = 4$, $n_2 = 15$, and consult Appendix G2. The conclusion is identical.

If the analysis of variance had shown a lack of uniformity between pencils, it must be apparent that a condition of general non-uniformity was indicated. The lack of uniformity between pencils might have been due to a large defection on the part of one or two pencils, while the others might have been relatively uniform. We could, of course, compare the mean of each pencil with each other one, as previously mentioned, and thus perhaps learn something more specific about the location of the non-uniformity.[13]

In the foregoing paragraphs the analysis of variance has been applied only to distributions in which $N_1 = N_2 = N_3 =$ etc. This is not a necessary condition and the formulae are but slightly altered if $N_1 \neq N_2 \neq N_3 \neq$ etc., as is shown in Appendix B, section XIII–2.

Further considerations of the analysis of variance will be found in the works of Fisher, Snedecor, and Tippett listed at the end of this chapter.

Criterion of Likelihood

Comparison of several σ's. Manufacturing concerns that are interested in maintaining uniformity of a product may find it necessary to compare, not merely two σ's or σ^2's. but a number of measures of variance (or uni-

---

[13] It is interesting to observe that, when there are two columns, the expression

$$\frac{\sum\limits_{1}^{m} \left[\sum\limits_{1}^{N_K} (Y - \overline{Y}_K)^2 \right]}{N - 2},$$ or the variance within columns, is equivalent to $\frac{N_1\sigma_1^2 + N_2\sigma_2^2}{N_1 + N_2 - 2}$, or

the estimate of σ_P based on two samples assumed to have been drawn from the same population, used in the preceding chapter. Furthermore, when only two categories are being considered, the significance of the variance between the two means (as given by the z or F test) is the same as the significance of the difference between the two means (as given by the t test). Both tests attempt to ascertain if the two samples were drawn from the same population.

formity) from samples of their product selected periodically or from each lot produced.

One method of comparing a series of sample σ's would be to compare σ_1 with σ_2, σ_1 with σ_3, σ_2 with σ_3, etc. Another procedure involves com-- paring all of the σ's at once[14] by means of a *criterion of likelihood*,

$$L = \frac{\sqrt[k]{\sigma_1^2 \times \sigma_2^2 \times \sigma_3^2 \times \cdots \times \sigma_k^2}}{\frac{1}{k}(\sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \cdots + \sigma_k^2)},$$

where k is the number of samples. When the samples are of varying numbers of items,

$$L = \frac{\sqrt[N']{(\sigma_1^2)^{N_1} \times (\sigma_2^2)^{N_2} \times (\sigma_3^2)^{N_3} \times \cdot \times (\sigma_k^2)^{N_k}}}{\frac{1}{N'}(N_1\sigma_1^2 + N_2\sigma_2^2 + N_3\sigma_3^2 + \cdots + N_k\sigma_k^2)}$$

where N_1, N_2, etc., are the number of items in the respective samples and $N' = N_1 + N_2 + N_3 + \cdots + N_K$.

The numerator is the geometric mean of the σ^2's, while the denominator is the arithmetic mean of the σ^2's. If there is any difference between the various σ^2's, the numerator will be smaller than the denominator. If all of the σ's are identical, there will be a condition of maximum uniformity and $L = 1$. The value of L has 0 as its lower limit, which represents a condition of maximum non-uniformity. This is a theoretical limit, which would not be approached in actual practice.

Let us compute the value of L for tests of strength of five pencils shown in Table 75. The first step is to compute the value of $\bar{\sigma}_1^2$, the variance of the first pencil. We use $\bar{\sigma}_1^2$ instead of σ_1^2 because the samples are of $N = 4$.

$$\begin{aligned}
\bar{\sigma}_1^2 &= \frac{\Sigma X_1^2}{N_1 - 1} - \frac{(\Sigma X_1)^2}{N_1(N_1 - 1)} \\
&= \frac{11.9420}{3} - \frac{(6.90)^2}{4(3)} \\
&= 3.980664 - 3.967500 \\
&= .01316.
\end{aligned}$$

In similar fashion we obtain:

$$\bar{\sigma}_2^2 = .05667, \qquad \bar{\sigma}_4^2 = .01930,$$
$$\bar{\sigma}_3^2 = .02787, \qquad \bar{\sigma}_5^2 = .01529.$$

[14] See J. Neyman and E. S. Pearson, "On the Problem of k Samples," Akademija Umiejetnosci, *Bulletin International de l'Académie Polonaise des Sciences et des Lettres,* Série A. Sciences Mathématiques, 1931, pp 460–481

We may now compute the coefficient of likelihood:

$$L = \frac{\sqrt[5]{.01316 \times .05667 \times .02787 \times .01930 \times .01529}}{\frac{1}{5}(.01316 + .05667 + .02787 + .01930 + .01529)}$$

Using logarithms to determine the fifth root of the products of the five variances, we have:

$$L = \frac{.02278}{.02646} = .86,$$

and it appears that there is uniformity among the variances of the five pencils tested.

Similar tests were made of pencils manufactured by five other companies. The product of one company showed $L = .92$, while at the other extreme another company's product gave $L = .30$.

It may be desirable to test the reliability of L. This is facilitated by referring to Nayer's tables, which are shown in Appendix H and which are constructed upon the assumption that the various samples have been drawn from a normal population. As noted before, $L = 1$ if all values of $\bar{\sigma}^2$ are identical. We therefore set up the hypothesis that $\bar{\sigma}_1^2 = \bar{\sigma}_2^2 = \cdots = \bar{\sigma}_k^2$, and ascertain whether a value of L such as that observed might occur by pure chance. It should be noted that the curve of the sampling distribution of L has its maximum ordinate at 1.0 and slopes downward (concave upward) to $L = 0$. It depends upon N (the number in a sample) and k. In the above instance we found $L = .86$, $N = 4$, $k = 5$. From Appendix H it is observed that $L = .491$ at the .05 level, and .370 at the .01 level. We conclude that the value obtained for L is not significantly less than 1; therefore our hypothesis is not impugned, and $L = .86$ for these pencils indicates real uniformity. Consider now the pencils manufactured by another firm which showed the least uniformity and for which $L = .30$. Here also $N = 4$, $k = 5$. Since the value of $L = .30$ is beyond the .01 level, the hypothesis of equality of variance of these pencils is very dubious, and we conclude that this product is clearly not uniform in regard to strength.

In this and the preceding section we have examined two measures of uniformity: (1) the analysis of variance which failed to indicate lack of uniformity of means from pencil to pencil, and (2) the coefficient of likelihood which has indicated uniformity of the variance within pencils. Observe that we have not considered whether or not this brand, designated "Pencil D," is stronger than some other make of pencil. As a matter of fact, it was significantly less strong than one other of six brands tested, data of which are not included in this volume. We do know, however,

that this brand possesses uniform strength, both within pencils and from pencil to pencil.

In these two chapters on reliability and significance we have discussed at some length the reliability of the arithmetic mean, of a percentage, and of the standard deviation, and also the significance of differences of such measures and of the variance between samples. In Chapter XI and in this chapter we made use of P values based upon the sampling distribution of χ^2. All statistical measures computed from samples are subject to sampling variation. In the later chapters dealing with correlation we shall discuss the reliability of various forms of the correlation coefficient.

Selected References

(NOTE: A number of references concerning χ^2 have been given at the end of Chapter XI.)

R. A. Fisher: *Statistical Methods for Research Workers* (Seventh Edition), Chapters VII, VIII; Oliver and Boyd, Edinburgh, 1938. Intraclass correlation and the analysis of variance.

R. A. Fisher and F. Yates: *Statistical Tables for Biological, Agricultural, and Medical Research;* Oliver and Boyd, Edinburgh, 1938.

A. B. Hill: *Principles of Medical Statistics*, Chapters IX, X; Lancet Limited, London, 1937. Use of χ^2 with applications to medical data.

F. C. Mills: *Statistical Methods Applied to Economics and Business* (Revised Edition), pages 473–474, 490–500; Henry Holt and Co., New York, 1938. Significance of difference between proportions and analysis of variance.

P. R. Rider: *An Introduction to Modern Statistical Methods*, pages 81–83, 117–119, 132–150; John Wiley and Sons, New York, 1939. Significance of difference between proportions and analysis of variance.

G. W. Snedecor: *Calculation and Interpretation of Analysis of Variance and Covariance;* Collegiate Press, Ames, Iowa, 1934.

G. W. Snedecor: *Statistical Methods Applied to Experiments in Agriculture and Biology*, Chapters 10, 11; Collegiate Press, Ames, Iowa, 1937. Analysis of variance.

L. H. C. Tippett: *The Methods of Statistics* (Second Edition), pages 84–86, 117–121, Chapter VI; Williams and Norgate, London, 1937. Reliability of and significance of difference between standard deviations (variances); criterion of likelihood, analysis of variance.

CHAPTER XIV

THE PROBLEM OF TIME SERIES

The Problem Stated

Economists are interested in two types of problems. First, it is possible to make an analysis of the situation which would logically exist if economic forces were in a state of equilibrium with no changes taking place. Starting with this ideal situation, we may then assume certain changes, and the new equilibrium which will finally result may be described. Thus, if the demand for a commodity were to increase a specified extent, how much would production expand, and how much would the price change? Such an analysis is usually referred to as static. A second type of analysis, often referred to as dynamic, which has engaged the attention of economists aims to explain what happens while the system is trying to reach equilibrium, rather than the situation that exists after this state is achieved.

It is but natural that economists with a scientific bent should devote much energy to the second type of analysis. Such analysis is concerned with the behavior of time series. Having undertaken investigations along this line, it is not surprising that tools would be developed especially suited to their purposes, and so we find that economics is largely responsible for the development of statistical methods for analyzing changes taking place over time. These methods are quite distinct from, though closely related to, frequency distribution analysis. Although the technique of time series analysis has been developed largely by economic statisticians, the study of time series is of interest to a wide range of people including businessmen, sociologists, biologists, doctors, and public health workers.

Characteristics of Time Series

Economists are not in complete agreement concerning the meaning of the various movements constituting time series, or the proper methods for their analysis. But even though the classification and the explanation of some movements be in doubt, certain characteristics of time series are

apparent upon very brief inspection. The movements which we shall consider in some detail are secular trend, cyclical, periodic, and irregular.

Secular trend. The gradual growth over a period of decades is perhaps the most striking characteristic of most industries. This is to be expected in a country such as the United States, with steadily growing population. But industrial growth is not completely accounted for by population changes; this is indicated by Chart 138. The natural sciences have been applied to industry and agriculture so as to increase their output enor-
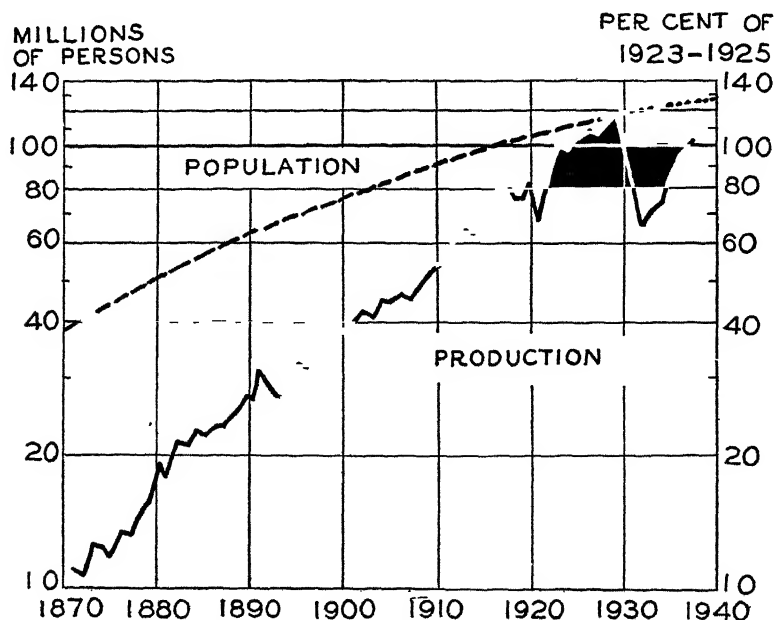
Chart 138. Index of Physical Production in the United States, 1870-1937, and Population of Continental United States, 1870-1940. (Production Index from Research Department of Federal Reserve Bank of New York. Population data from United States Bureau of the Census, *Abstract of the Fifteenth Census of the United States, 1930*, p. 9. 1940 population estimate from Table 88.)

mously. Keeping pace with these technological changes and induced by them have been changes in business organization and methods. The growth of the corporation has permitted the accumulation of sufficient capital for specialization and mass production, while scientific management and personnel management have found their way into many organizations.

In addition to these factors which affect the growth of all industries, we find that some industries wax or wane because of changes in demand New commodities may attract favor and replace old ones fulfilling a similar need, as the automobile did the horse and buggy. No doubt the

automobile has also drawn purchasing power away from commodities catering to quite different desires. Demand may also be drawn off, though desire may be undiminished, through the appearance of a less attractive, though cheaper, substitute. Thus rayon is partially replacing silk. More spectacular is the development of the railroads, forcing into obsolescence many canals, only to have their traffic more recently diverted by competition of trucks, buses, and airplanes. Although physical production in the United States as a whole seems to have been increasing, until very recently, at a fairly constant rate, many industries, such as pig iron production, shown with a logarithmic vertical scale in Section A of Chart 139, seem to be characterized by a rather steadily declining rate of growth. This may be due to a combination of a number of reasons. Improvements in the productive process are rapid at first, but as time goes on it is possible that further improvements have less and less effect upon the output. Again, growth may be retarded by the increasing difficulty of obtaining supplies, such as minerals, for mining becomes more and more difficult as the better ores are exhausted. Further, during the period while difficulty is experienced in keeping up with demand, profits will be high and it will be easy to expand productive equipment. But after a while recourse must be had to the open market for funds. Eventually the point will be reached where funds, though forthcoming in large amounts, are small relative to the size of the business. Finally, as consumer desire in old markets becomes more nearly satisfied, relative to that for other commodities, it becomes increasingly difficult to entice buyers from competing products, and new markets may not appear.

Many authorities think that not only does the rate of growth decrease, but eventually further expansion will be physically impossible. Raymond B. Prescott has characterized the tendency we have described as a "law of growth,"[1] which applies to all industries. This law embraces four stages: (1) period of experimentation during which the amount of growth is small; (2) period of growth into the social fabric; (3) period during which growth is retarded as saturation point is approached; (4) period of stability. Section B of Chart 139 (on arithmetic paper) indicates that the trend of pig iron production answers this description also. Although it is quite apparent by inspection of Sections A and B of Chart 139 that a trend with a decreasing *rate* of increase may be one which varies in *amount* of growth, as described by Prescott, nevertheless the former type need not decline in amount of growth during its latter stages, nor need a curve which in its early stages is growing by increasing *amounts* be also one which is increasing at a *decreasing rate*.

---

[1] "Law of Growth in Forecasting Demand," by Raymond B. Prescott. *Journal of the American Statistical Association*, December 1922, Vol. XVIII, pp. 471–479
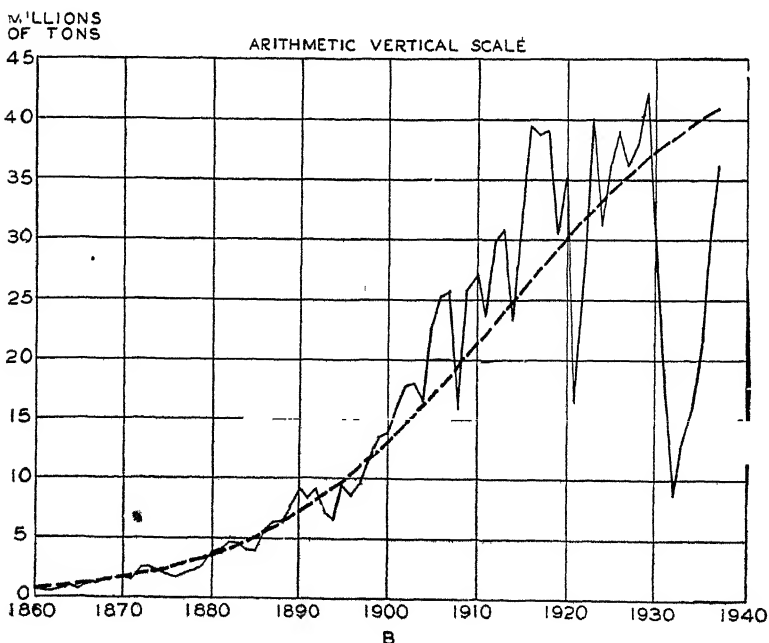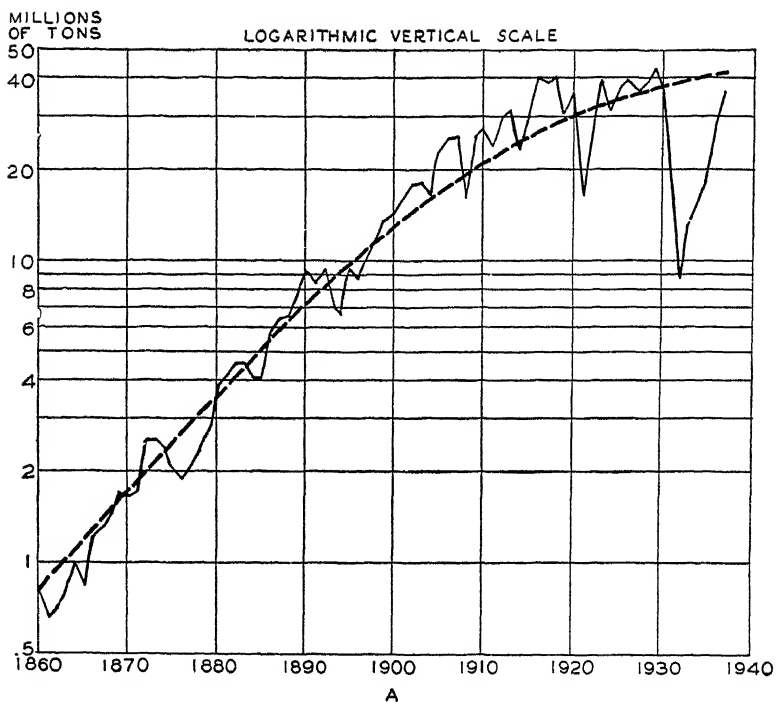
Chart 139. Pig Iron Production in the United States, 1860-1937, and Secular Trend: A. Logarithmic Vertical Scale; B. Arithmetic Vertical Scale. [1860–1918 data from Interior Department, United States Geological Survey, *Mineral Resources of the United States*, 1918, Part I (Metals), p. 566. 1919–1937 are annual totals of monthly production; source, *The Iron Age*, as quoted in Standard Statistics Co., *Standard Trade and Securities*, *Basic Statistics*, Vol. 80, June 5, 1936, p. G-5, and *Current Statistics*, Vol. 90, October 14, 1938, p. 23.]

There is a difference of opinion whether price level changes may be said to have a trend. Probably, however, it is useful to study both long time and short time changes in price. The former are due to such factors as variations in the stock of gold and in coinage laws, and to changes in business methods, in personal monetary habits, and in banking technique. The latter are subject to quite different influences. Thus, although trends in price are the result of factors other than those affecting production trends, and therefore behave differently from them, they are both long time changes and both are norms around which other movements fluctuate.

The statistical problem is, first, to decide what type of trend fits the data closely and is a logical description of them, and, second, to fit the trend of the type decided upon. Such a trend is not only an expression of normal[2] tendencies; it also provides a base from which to measure deviations.

Cyclical movements. Business cycles are a type of fluctuation lasting longer than one year that tend to recur with a measure of regularity in economies organized on a business basis. Such movements are called cyclical rather than periodic because they do not occur with complete regularity as to duration. On the other hand, they are cyclical rather than random movements because the position of business in the cycle is affected by the position of business in recent months and, in turn, affects business in the more immediate future. In other words, the transition from a low point to a high point, or vice versa, is a progressive development. Cycles appear to operate somewhat on the principle of a pendulum. Just as a pendulum is pulled by gravity toward a vertical position, but tends constantly to move past its position of equilibrium, so it is said that business is drawn toward an equilibrium by the forces of demand and supply, and so also do the errors in one direction tend to progress into errors in the opposite direction. Such an explanation of business cycles is known as the "self-generative theory," usually associated with the name of Wesley C. Mitchell. But just as the mechanism impelling a pendulum must be wound up occasionally, so it is possible that economic activity would attain equilibrium were it not for other propulsions of varying degrees of intensity. It is possible to speak of cycles in general business or of cycles in particular industries, such as residential construction, cattle raising, or textile production. Occasionally cycles in a specific industry appear to be inherently periodic, as in the case of the two-year cycles in rayon consumption. In any event they are modified by the position of

---

[2] The reader should not confuse the statistical meaning of the word "normal," which is a sort of average, with another meaning of the same word, as used in theoretical economic analysis, to designate the situation which would exist if the economic system were in equilibrium.
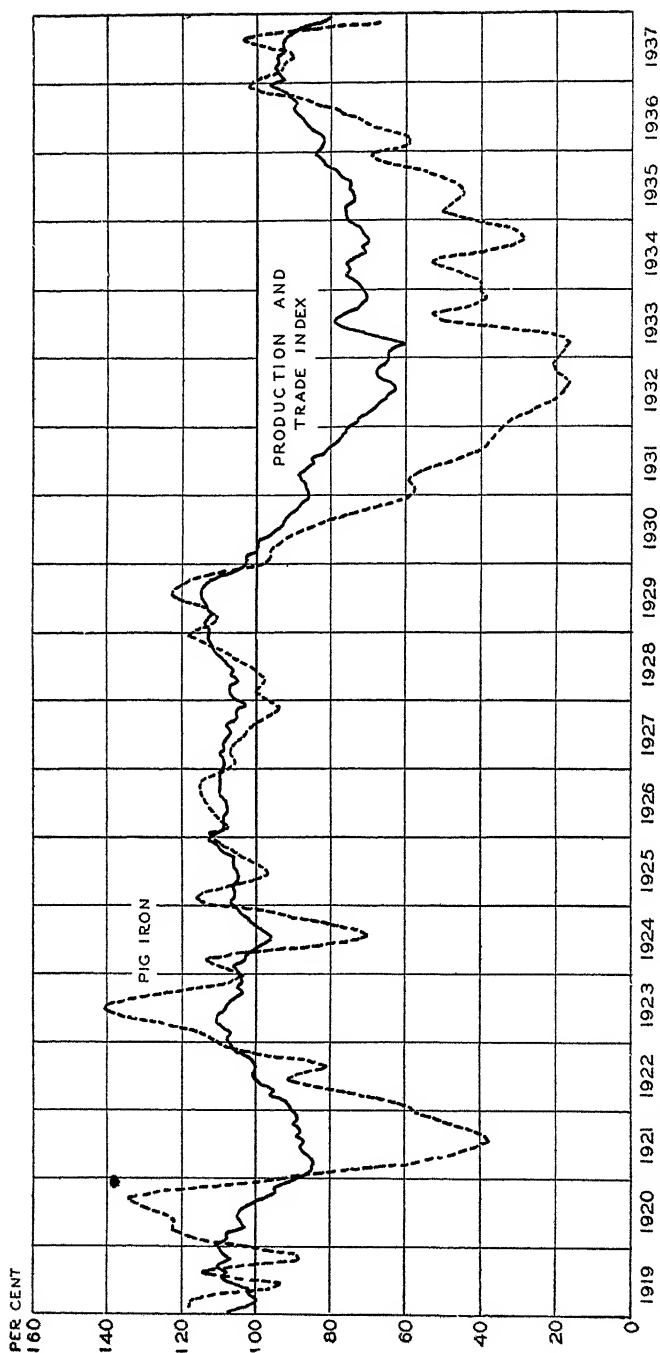
Chart 140. Cyclical Movements of Pig Iron Production, and Federal Reserve Bank of New York Index of Production and Trade, 1919-1937. (For source of pig iron data see Chart 139; trend and seasonal influences have been removed and the adjusted data smoothed slightly to reduce minor fluctuations. Production and Trade Index is from Federal Reserve Bank of New York)

the general cycle. However, since all industries are so interdependent, a revival or recession in a key industry or group of industries soon transmits its effect to the other branches of activity.

It appears that cyclical movements of general activity could be generated by a concurrence of the same cyclical phase in the activity of several important industries; or they might be generated by interferences from outside the business world. These interferences might be occasional events of considerable magnitude, such as a war, a discovery, unusual weather, or some political event; or they might be the simultaneous occurrence of several minor events, each reinforcing the effect of the other.

The rough regularity of cycles may possibly be explained by the periodicity of certain of the extraneous events which, some authorities believe, are in part responsible. Cycles in weather have been suggested. It is more likely, however, that what regularity can be observed is due to the fairly constant length of time it takes the business world to respond to stimuli. For instance, the time it takes for erecting a building or for foreclosing a mortgage, or even to decide to go into bankruptcy, is not utterly irregular. Perhaps greater regularity would be observable were it not for the irregularity of accidental occurrences.

There are some who reject the concept of self-generating cycles, believing that cycles are brought about largely by external influences. Even these, however, are interested in observing whether production and consumption are increasing or decreasing, and especially in taking practical measures for stabilization. Whether self-generated or entirely caused by non-business occurrences, it can be seen, from Chart 140, that pig iron production has experienced recurring depression every three or four years since 1919. Furthermore, the variations are very similar to those of the total volume of trade. It must, of course, be recognized that pig iron production is one of a large number of series represented in the total volume of trade. The greater amplitude of fluctuation is partly a characteristic of pig iron, a producer's good, and is partly due to the fact that an index composed of several series whose turning points occur at slightly different periods of time always cancels out some of the amplitude of the constituent series by the averaging process. Although the average length of time from depression to depression of pig iron as shown by this chart is about 45 months, there is considerable variability in the duration of the different cycles. Also, it might be noted that there is considerable difficulty in deciding just what is a cycle. Were, for instance, the slight recessions in 1925 and 1934 cycles or large irregular occurrences? If cycles, the average length of cycle was much less than 45 months.

Periodic movements. As distinguished from cyclical movements, which have only rough regularity, many time series have variations which repeat

themselves with remarkable similarity at regular intervals. Chart 141 shows the variation in the number of automobile injuries in New York City during different hours of the day. An example of a type of movement that repeats itself each week is the circulation of books in the reserve room of a university library. (See Chart 173 in Chapter XVII, dealing with periodic movements.) A still longer periodicity is the intra-month variety. Thus bank debits have a tendency to reach a peak shortly after the first of each month.

NUMBER OF
ACCIDENTS
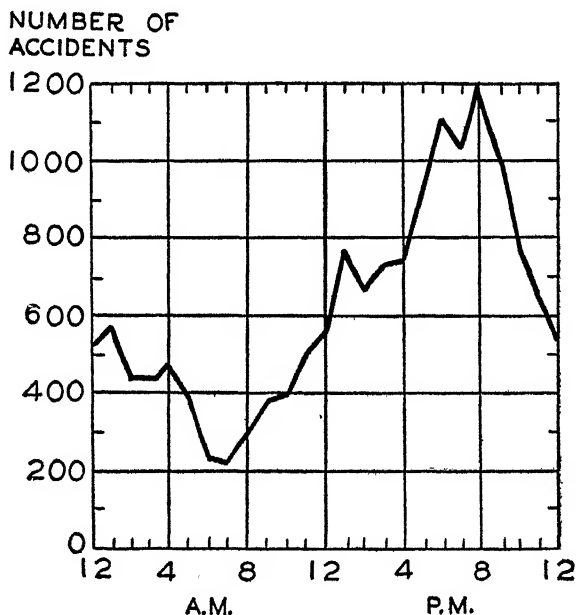
Chart 141. Average Hourly Number of Automobile Injuries that Occurred in New York City During the Six-Month Period From January to June, 1937. (Traced from chart appearing in The New York *Times*, September 1, 1937.)

The type of periodic movement which has engaged much of the attention of economists, however, is that which has a period of one year, and is commonly known by the term, *seasonal variation*. (See Chart 142.) Climatic conditions, such as variations in rainfall, snow and ice, sunshine, humidity, heat, and wind produce variations in demand which often reflect themselves in variations in production, and also directly affect production in such occupations as agriculture and building construction. Social conventions also have their influence, the Christmas trade being a notable instance. To some extent holidays are not entirely independent of seasonal factors. Easter and Thanksgiving owe their origin at least in part to weather conditions. Also, it might be noted that man's propensity for

ostentation leads him to change his style of car or coat each year at the proper season.   These style changes greatly accentuate the seasonal variations for which nature is primarily responsible

For statistical analysis it may be desirable to calculate seasonal variations on either a monthly or a weekly basis.   On both of these bases, typical variations in steel mill capacity may be seen in Chart 142.

It is worth observing that a seasonal pattern may change either gradually or suddenly with the passage of years.   Thus, a study of the seasonal
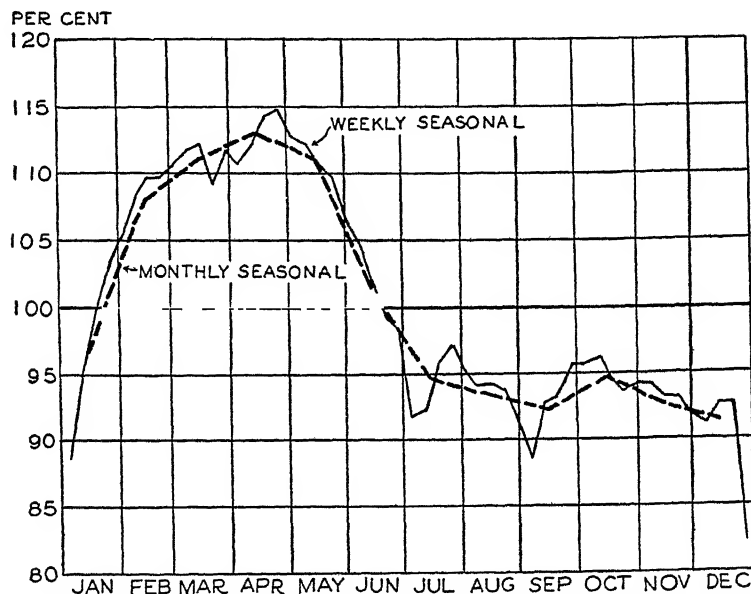
Chart 142.   Typical Monthly and Weekly Variations During a Year in Per Cent of Capacity of Steel Mill Activity.   (Computed from *Wall Street Journal* data as supplied by Standard Statistics Co.)

curve of Chart 143A indicates that while pig iron is now typically at a low point during the winter months, this has not always been the case. On the other hand, Chart 187, page 507, indicates that the Christmas trade of department stores has become increasingly more important.   In the automobile industry, on the other hand, there was a sudden shift in 1935 during the time when new models were introduced, with a consequent effect on production schedules.   This may be clearly seen from Chart 192, page 517.   Some series retain the same general pattern but change in intensity gradually or irregularly from year to year.   This is particularly true of agricultural series.   See, for instance, Chart 193, page 520, dealing with receipts of sheep and lambs at primary markets.   Still other series may

retain a constant seasonal pattern, but exhibit peaks and troughs at different months in different years, because of early or late seasons.

Irregular variations. There are other variations which are not covered by the above classification. No theory seeks to explain these variations, and they may be considered as accidental from the point of view of the theorist. They may be minor fluctuations, perhaps of a random nature, too small to be worth considering individually, or they may be important episodic events, such as wars, earthquakes, or general strikes. As suggested above, these episodes may be so important as to generate, or assist in generating, a cyclical fluctuation, and occasionally it may be difficult to distinguish an episode from a cycle

A Graphic Illustration

Perhaps the nature of the movements will be more clearly understood if they can be seen graphically in one chart. The different movements for
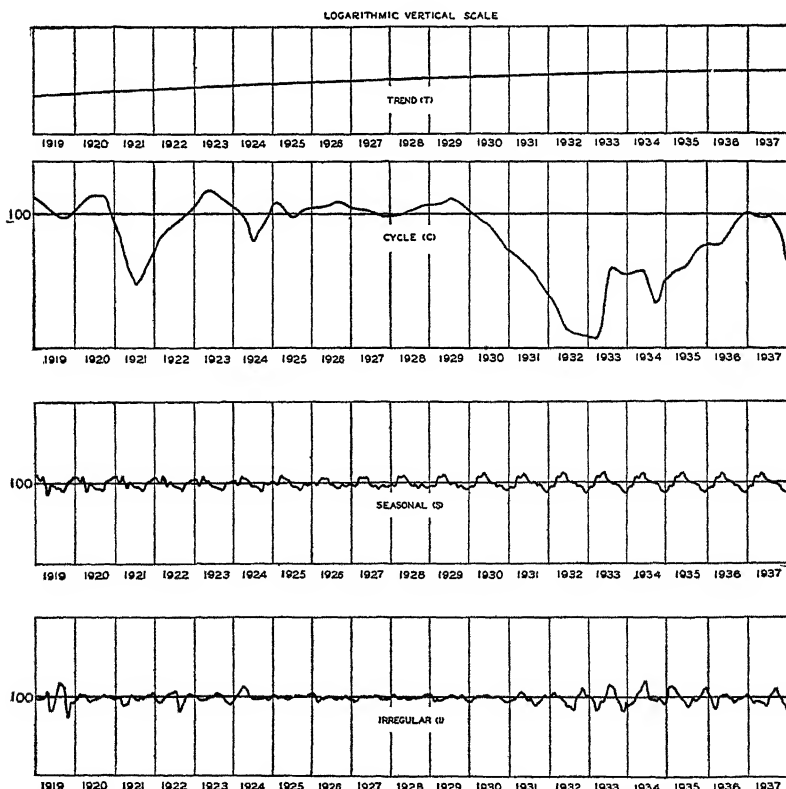
Chart 143A. Graphic Analysis of Variations in Pig Iron Production in the United States, 1919-1937. (For source of data from which these variations were estimated see Chart 139.)

pig iron are shown separately by the different curves of Chart 143A. The upper curve represents the secular trend and will be recognized as a fragment of Chart 139A. It is not nearly so steep as it was in earlier years. The curve immediately below is an estimate of the cyclical movements of pig iron. The very large amplitude of these variations is immediately apparent. Of much smaller amplitude is the seasonal variation, shown as the third curve in this section. As can be seen, the seasonal pattern is not stable but is gradually changing. The irregular movements are indicated by the curve at the bottom. Their amplitude is, in general, rather large in comparison with that of the seasonal variation, although they seem quite modest during the middle portion of the period covered. It should not, of course, be assumed that the relative amplitude of the different movements is the same for other series as it is for pig iron production. Each series is characterized differently in this respect.
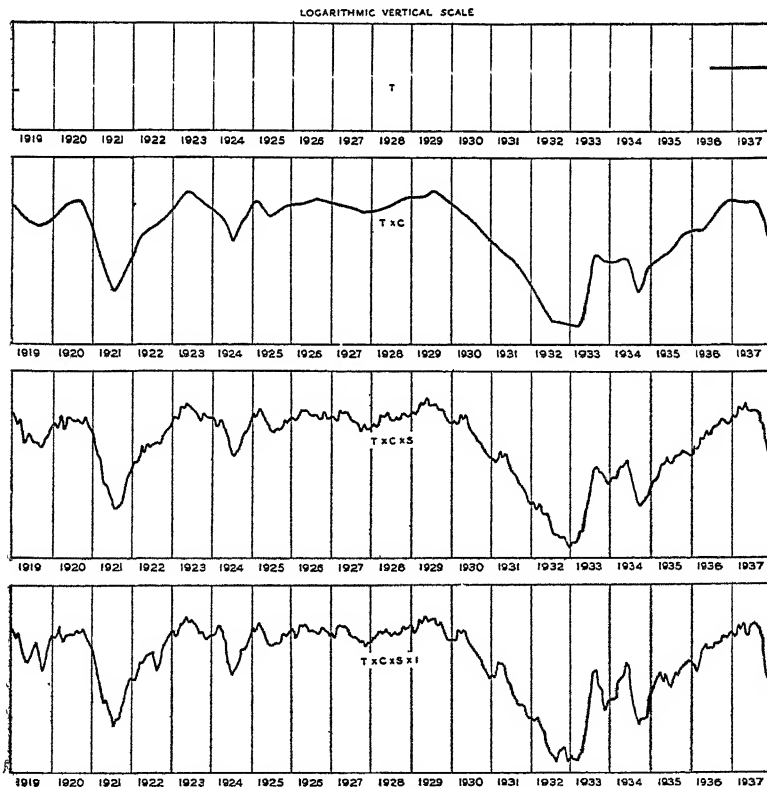
Chart 143B. Graphic Synthesis of Variations in Pig Iron Production in the United States, 1919-1937. (For source of data from which these variations were estimated see Chart 139.)

Turning to Chart 143B, we find a progressive synthesis of the different elements beginning with the trend and ending with the original data. The chart is semi-logarithmic, so that any curve in Chart 143B can be obtained by graphically adding the different curves of Chart 143A, beginning at the top and ending with the one to the left of the one being synthesized. (Or, any curve of Chart 143B can be obtained by a graphic summation of two curves, the one immediately above it and the one immediately to its left in Chart 143A.) It is apparent also that, if we should add for any month the logarithms of the values represented by every curve of Chart 143A, we would obtain the logarithm of the corresponding monthly value of the original data represented by the $T \times C \times S \times I$ line of Chart 143B. This is equivalent to saying that the original data are the product of trend, cycle, seasonal, and irregular movements,[3] or

$$\text{Original data} = T \times C \times S \times I.$$

TABLE 77

PIG IRON PRODUCTION AS SYNTHESIZED FROM ITS ESTIMATED COMPONENT ELEMENTS, 1936

Month	Secular trend in millions of long tons T	Cyclical movements as ratios C	Seasonal movements as ratios S	Irregular movements as ratios I	Actual production in millions of long tons $T \times C \times S \times I$
January . .	3,358	.650	940	9869	2,025
February .	3,361	.650	950	.8788	1,824
March . . .	3,364	.650	1 080	.8638	2,040
April. .	3,367	650	1 080	1 0171	2,404
May. . ..	3,370	.667	1 150	1 0244	2,648
June.	3,374	.701	1.050	1.0413	2,586
July...... .	3,377	.751	1 025	.9979	2,594
August.. .	3,380	.811	1 005	.9841	2,711
September	3,383	874	.980	.9421	2,730
October. ...	3,386	929	.970	.9805	2,992
November ..	3,389	.970	.900	.9962	2,947
December. .	3,392	1 000	870	1.0555	3,115

Source T C S, and I computed from actual production   Actual production from 1919 to date, Stand ard Statistics Company, I. c, Standard Trade and Securities, Basic Series No. 80, June 5, 1936, p. G-5, and *Current Statistics*, Vol 89, August 1938, p 23

[3] It is also possible to consider that

$$\text{Original data} = T + C + S + I.$$

That is not, however, such a generally useful concept, since C, S, and I tend to remain about constant in magnitude relative to trend. This makes it possible to compute a seasonal index which remains uniform over a period of years; or to compare the percentage fluctuations of cyclical variations. But there are series which give better results when seasonal is considered as constant in absolute, rather than relative, magnitude See pages 525–527 for further discussion of this point.

Charts 143A and 143B illustrate the various individual movements of which pig iron production is composed, as well as a number of combinations of movements.   Sometimes we wish to study the trend alone, sometimes the seasonal, frequently the combination of trend and cycle, and possibly most often of all the cyclical movements.   Inspection of Chart 143B and of Table 77, in which the movements of pig iron production are synthesized from the elements of the series, should also explain the logic of the usual method of analyzing time series to obtain cyclical movements as a final product.   When analyzing, we first estimate by statistical
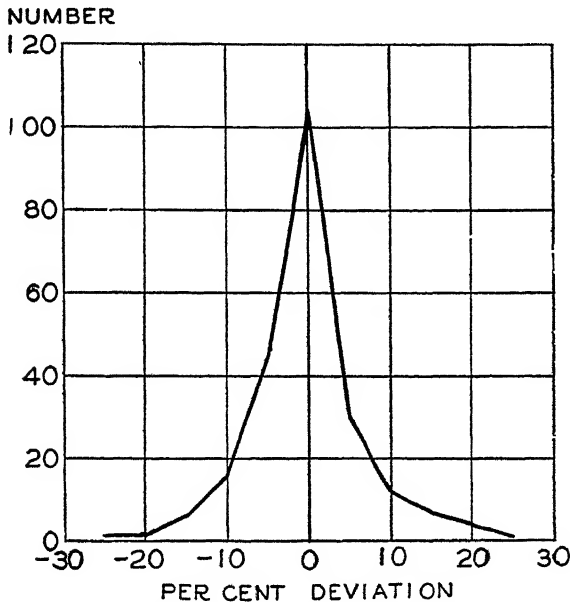
Chart 144.   Irregular Variations in Pig Iron Production Classified by Magnitude of Deviation.   (This frequency distribution was obtained by grouping the irregular deviations of Chart 143A.)

methods: (1) trend, and (2) seasonal variation; then the data are divided through successively by the trend values and the seasonal values (or vice versa); finally, irregular variations are smoothed out, leaving only the cycles.   These processes will be explained in subsequent chapters.

In Chart 144 the irregular movements are shown as a frequency curve. The distribution plainly is leptokurtic.   If the irregular movements had been of a random character, we would have expected a normal distribution, but in addition to minor fluctuations we have others that are episodic in character and whose effects are cumulative over several months.   Note that in Chart 143A the irregular movements sometimes remain on one

side of the 100 per cent line for more than six months in succession. Actually, the interdependence of successive observations in a time series destroys the random character of any of its movements, regardless of the character of the factors causing such movements.

The different movements which we have discussed—trend, cycles, seasonal movements, and irregular fluctuations—are of varying importance in different series. At the present time the outstanding feature of such a series as rayon consumption is its steep secular trend. Durable goods, such as pig iron, fluctuate tremendously with the course of business cycles; still others, such as department store sales, do not show a steep trend or pronounced cycles, but exhibit intense seasonal fluctuations. Individual series, particularly those with a narrow or specialized market, often are quite irregular in their variations, but broad group indexes present smoother curves when seasonal movements have been removed.

Another View of Time Series Movements

It is probably an over-simplification to say that there are only four types of movements discernible in time series. Some analysts would say that time series are composed of wave-like movements of various lengths superimposed upon each other (including, of course, those already discussed). Some of these movements have been the subject of investigation by economists. Thus Kondratieff[4] has discovered "long cycles" lasting roughly 50 years, running through many series and in a number of countries. He lists these waves as follows:

Wave	Low	High
I	1780–1790	1810–1817
II	1844–1851	1870–1875
III	1890–1896	1914–1920

He finds that prices fall and agriculture suffers during the decline, and also that scientific discoveries are made during this period. At the beginning of the upswing, colonies are acquired and new sources of gold found; during the upswing, scientific discoveries are applied, and there are wars and revolutions. These swings are held to be cyclical in character since the factors with which they are associated are, in part, at least, the result of the preceding phase of the long cycle. Thus falling prices, which are associated with the downswing, lead to a search for new technical methods to lower the cost of production, and for the increasingly valuable commodity, gold.

Kuznets has extensively studied another type of fluctuation, intermediate in length between Kondratieff's long cycles and the ordinary business cycle. These waves he calls "secondary trends."[5] The heavy solid line of Chart 145 show~

---

[4] See "The Long Waves in Economic Life," by N D. Kondratieff, *The Review of Economic Statistics*, November 1935, pp. 105–115 (translated by W. F. Stolper)

[5] See Simon S. Kuznets, *Secular Movements in Production and Prices*, Houghton Mifflin Company, Boston, 1930.
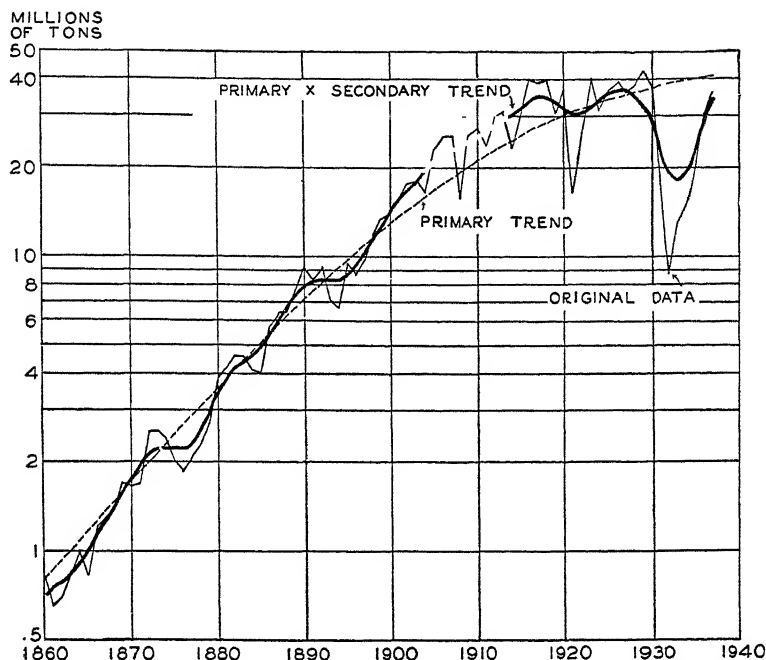
Chart 145.   Pig Iron Production, Primary Trend, and Primary × Secondary Trend by Years, 1860-1937.   (For source of data see Chart 139.)
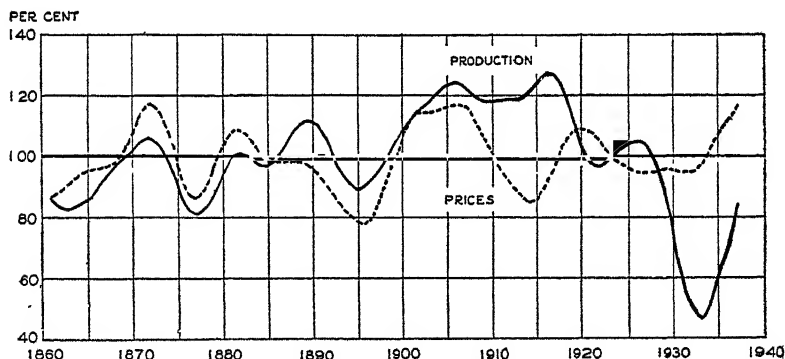
Chart 146.   Secondary Trends in Production and Prices of Pig Iron in the United States, 1860-1937.   (Secondary Trends are b.nomially weighted 15-year moving averages; 1860 and 1931–1937 trend values are estimates.   For source of production data see Chart 139.   Pig iron prices, 1860–1913, are from Simon Kuznets, *Secular Movements in Production and Prices*, Houghton Mifflin Co., Boston, 1930, pp. 364–365; for the years 1862–1868 the prices have been reduced to a gold basis.   For 1914–1937 basic prices have been spliced to Kuznets' No. 1 foundry prices.   These are from *The Iron Age* as quoted by Standard Statistics Co., *Standard Trade and Securities, Basic Statistics*, Vol. 80. June 5, 1936, p. G-6. and *Current Statistics*, Vol. 90, October 14, 1938, p. 23.)

such fluctuations around the primary trend line, while in Chart 146 they are shown as percentages of primary trend. It may be that these secondary trends are a disequilibrium in production resulting mainly from a disequilibrium in prices It is commonly believed that rising prices stimulate business. This is largely because many of the businessman's expenses lag in changes behind his selling price. If his selling price merely kept pace with wages and other costs, his profits would not increase, but when prices of a particular commodity are above normal, then it is profitable to expand. The initial stimulant of abnormal prices tends to be stretched out over more than one cycle. Price movements themselves may be initiated by such factors as variations in the volume of gold production or changes in banking technique, which phenomena are often gradual developments that are in progress for a considerable period of time. The plausibility of the theory is indicated by comparing the secondary trends in production and prices of pig iron shown in Chart 146. To make secondary trends fall into the self-generating category, it is necessary to show that the factors generating the price rise are attributable to the preceding downswing of prices.

Long cycles, secondary trends, and business cycles are each a combination of outside influences and business responses; but in the first of the three it may be that the former predominates, while in the third the self-generative aspect may be most predominant. These three movements merge into one another and in actual practice it is difficult to separate them. Should, for instance, the depression following 1929 be regarded as a very severe depression, or was it the coincidence of a business depression with the low point of a secondary trend? Possibly, future years will establish that a new primary trend is appropriate for the years following, say, 1931. Furthermore, it is difficult to say when a short wave-like movement is a cycle and not an accidental variation. This cannot be decided by the criterion of cause, for many cycles are partly caused by forces external to business. More reasonable is to consider as cycles only those movements which extend over a wide area of our business life. But the question immediately rises: How wide? Was, for instance, the wave that rose in the spring of 1933 and reached a trough in the fall of 1934 a cycle? The answer is, of course, subjective, and seems to boil down to saying that any wave may be considered a cycle if we wish to study its cyclical characteristics. As a guide to economists and statisticians, Willard L. Thorp has studied the economic history of various countries, and on the basis of all evidence, statistical and otherwise, has subdivided the period since 1790 into sub-periods according to the different phases of the business cycle. His findings are published in *Business Annals*.[6] More difficult of identification than the cycles of general business are those pertaining to a particular industry. Attention has already been called to the difficulty of identifying cycles in pig iron production.

Preliminary Treatment of Data

Some variations in time series may be due to the terms in which they are stated. Before attempting to isolate for purposes of study the move-

---

[6] Published in 1926 by the National Bureau of Business Research, New York. This study is kept up to date by the National Bureau, with results published in occasional bulletins.

ments which have been described, it may be well to restate the data in more significant terms.

Calendar variation. Usually, though not always, there are 365 days in a year. Although there are 12 months in each year, they vary in length from 28 to 31 days. To make matters more complicated, the different months do not start on the same day of the week, nor does the same month in successive years so start. Another difficulty is the matter of holidays Not only do the number of Saturdays and Sundays vary as between months, but February, with 28 or 29 days, has Washington's birthday and Lincoln's birthday, while March, with 31 days, usually includes no holidays. Even more confusing is the way Easter fluctuates between March and April. Thus do the different months vary extremely as to the number of working days.

Although it seems impossible to divide the year into quarters containing the same number of whole weeks, nevertheless some business firms have tried to minimize the difficulty. A few firms keep records by 4-week periods. There are 13 such periods in a year, but quarterly data cannot be kept by this system. A few others keep records by quarters, each quarter being composed of three months—the first two months of four weeks each and the third of five weeks. Of course, neither of these plans is satisfactory so long as the first of a given calendar month may occur in either of two artificial months. And under any plan the unsystematic occurrence of holidays results in a different number of working days in successive artificial months. Movements have been launched to change the calendar to remedy these defects. One plan suggests identical quarters; each quarter would contain, not identical months, but three monthly patterns of thirty or thirty-one days each, these three patterns being repeated so as to occur four times a year. An extra day, however, known as Year Day, would occur at the middle of the year.

But until people can be persuaded to change their established customs sufficiently to change their calendar, the statistician is confronted with the problem of adjusting for calendar variation. The method is very simple. Using electric power production in 1930 as an illustration, and assuming that adjustment for the number of calendar days is sufficient, we may divide the values of each month by the number of days in that month, thus expressing the data as millions of kilowatt hours per calendar day. This procedure is followed in Table 78, the results being recorded in column 4. The data so adjusted will be used in the chapter on periodic movements. If it is desired to retain the figures in their original magnitude, the figures as shown in column 4 must be multiplied by 365 ÷ 12 = 30.4167, the average number of days in a month. Or, the same result can be obtained by dividing the original data by the ratio of the actual

to the average number of days in each month. These ratios are shown in column 5 of Table 78, and the results in column 6. These data are now spoken of as production "adjusted" for the number of calendar days, or for calendar-day variation. That the data so adjusted are less irregular is easily seen from Chart 147.

TABLE 78

Adjustment of Electric Power Production for Number of Calendar Days in Each Month, 1930

(Millions of kilowatt hours)

Month (1)	Actual production (2)	Calendar days (3)	Production per calendar day [Col 2 ÷ Col 3] (4)	Ratio of actual to average calendar days* (5)	Product in terms of average calendar days [Col 2 ÷ Col 5] (6)
January	8,663	31	279 5	1 01918	8,500
February	7,627	28	272 4	92055	8,285
March .	8,187	31	264 1	1 01918	8,033
April ..	8,019	30	267.3	98630	8,130
May ..	8,064	31	260 1	1 01918	7,912
June .	7,784	30	259 5	.98630	7,892
July. .	7,899	31	254 8	1 01918	7,750
August .	7,906	31	255 0	1 01918	7,757
September .	7,792	30	259 7	.98630	7,900
October .	8,195	31	264 4	1.01918	8,041
November .	7,693	30	256.4	98630	7,800
December.	8,108	31	261 5	1.01918	7,955

* Col. in (3) divided by 30 4167 = 305 ÷ 12
Source United States Department of Commerce, *Survey of Current Business*, 1936 Supplement, p 85

If, however, it is desired to adjust for the number of working days, a little more labor is required. The procedure is as follows:

(1) Ascertain the schedule of holidays appropriate to the industry. This, of course, varies with industries and localities.[7]

(2) Count the number of Sundays in each month of each year. If Saturday is not a working day, the number of Saturdays must be counted also. If a half holiday, Saturday is given a weight of one-half.

(3) Count the number of holidays in each month of each year. Perhaps some holidays will be given half weight.

(4) Add the number of holidays and the number of Sundays (and perhaps Saturdays) for each month. For many industries an extra holiday must be added if a regular holiday occurs on Sunday.

(5) Obtain the number of working days for each month by subtracting

---

[7] For a schedule of holidays by states, see "Legal Holidays in the United States, 1936,' *Monthly Labor Review* Vol. 34, No. 5, November 1936, pp. 1193–1196.

the number of holidays, Sundays (and perhaps Saturdays) from the number of calendar days.

(6)   Divide the original data by the number of working days each month; or adjust for working days by dividing by the ratio of the actual to the average number of working days.

The laborious part of this procedure consists, in large degree, in the necessity of computing the number of working days.   To facilitate discovering the number of Saturdays and Sundays in different months, as well as the months in which principal holidays occur on a Saturday or a Sunday, a flexible calendar has been included as Appendix K.
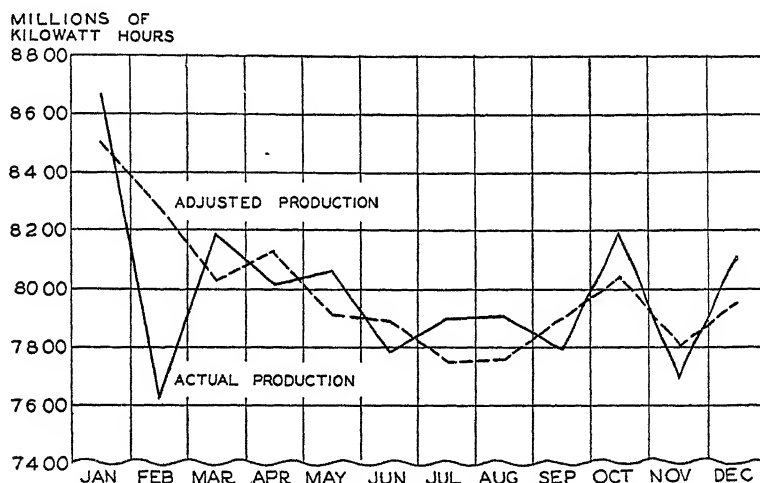
Chart 147.   Electric Power Production, by Months, 1930, Before and After Adjustment for Varying Number of Calendar Days in a Month.   (Data of Table 78.)

Not all time series require adjustment for calendar variation.   Clearly it would be spurious to do so for salary expenses of most corporations, since salaries of executives are usually constant from month to month. But for data requiring such adjustment it is frequently a difficult statistical problem to decide whether to adjust for working days, or merely for calendar days.   For some commodities it can logically be maintained that holidays within a month, far from decreasing consumer purchases during that month, may actually increase them.   If the holiday occurs on the last day of the month and the stores are closed, however, it might decrease sales.   In organizations which receive orders through the mail from a considerable distance, sales may be decreased as well by holidays occurring during the last few days of the preceding month.   Just what is the logical adjustment to make is often very difficult to determine and requires fa-

miliarity with the business or industry in question. In case of doubt it is always possible to determine experimentally what method gives the smoothest results after the adjustment is made. Such a test provides no conclusive evidence but is only presumptive. Frequently, a separate adjustment should be made for Easter, as explained on pages 509–515.

Population changes. Since one element in primary trend is population change, it may be worth while to adopt the old military axiom of divide and conquer by expressing the data on a per capita basis. Or, the data may be adjusted for population changes by dividing by population figures relative to some base. This is done for Barron's Index of Production and Trade in Table 91, page 415. The mechanical process is to divide the original data by the population figures. Naturally the remaining trend has quite a different character. Frequently it is simpler than before.

TABLE 79

Factory Average Hourly Earnings (25 Industries) and Cost of Living, 1929–1937

Year (1)	Hourly wage (cents) (2)	Cost of living (1929 = 100) (3)	Hourly real wage (cents) [Col 2 ÷ Col. 3] (4)
1929	59.0	100.0	59.0
1930	58 9	96 6	61.0
1931	56 4	87 1	64 8
1932	49 8	77.8	64 0
1933	49.1	74 8	65 6
1934	58.1	79.3	72 3
1935	60 0	82.5	72 7
1936	61 7	84 6	72.9
1937	69 3	88 4	78 4

Source· *Survey of Current Business*, 1936 Supplement, pp 11, 41, March 1937, pp 23, 31, March 1938, pp 63, 71   Cost of living index base has been shifted from 1923 to 1929 for this illustration

Price changes. Since economists are usually interested in physical volume changes rather than value changes, it is often necessary to convert the figures into this form. The process is generally called *deflating*. It consists in dividing the value figures, period by period, by the appropriate price index figures, as illustrated by Table 79, which shows figures on hourly wages and cost of living. The logic of this procedure is simple. Since price times quantity equals value, quantity must equal value divided by price. The hourly real wage column may also be referred to as hourly wages in terms of 1929. The table shows that hourly wages of employed persons did not fall as rapidly as cost of living during the depression. Hourly real wages, therefore, have increased, as can be seen from Chart 148. This, of course, does not imply that the employed wage

earner's position improved during the depression, since the number of hours worked per week was frequently shortened.   It does indicate, however, that one hour's labor enables the worker to command more commodities and services than formerly.   It is to be noticed that a cost of living index (rather than an index of the general price level, or a wholesale commodity price index) is used as a deflator.   Unless a deflator is used that pertains to the data being deflated, a satisfactory measure of physical volume cannot be obtained.

Securing comparability.   Statisticians for trade associations experience considerable difficulty in obtaining prompt reports from all members.   For
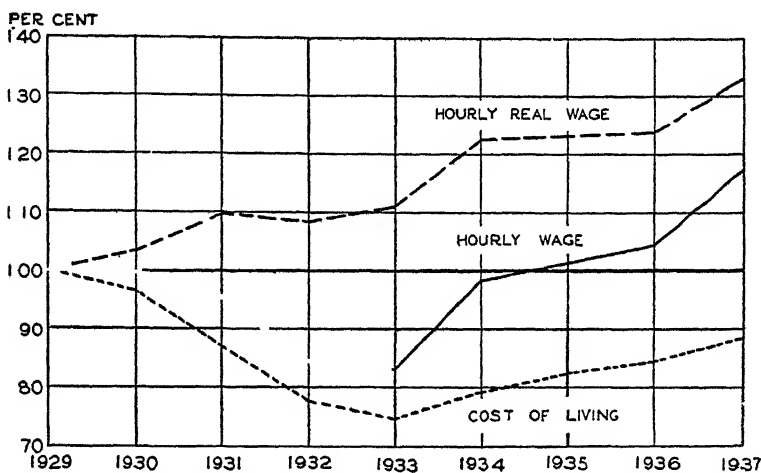
Chart 148.   Nominal Hourly Wages, Cost of Living, and Real Hourly Wages, 1929–1937.   (Data are from Table 79, but wage data have been expressed as percentages of 1929 to facilitate comparison.)

instance, 93 firms might report on time one month and 96 the next, the latter not necessarily, however, including all the 93 firms.   To be strictly accurate, a new time series should be constructed each month *for the entire period* including all of, and only, those firms which reported promptly for the month in question.   Thus, a complete time series one month would be computed for the 93 firms, and the next month for 96.   This is a very laborious procedure.   An easier procedure is to make a preliminary estimate by computing the percentage of the preceding period for only those firms which reported promptly in the two consecutive months, and to multiply the figure for the preceding month (which now includes all firms) by this percentage. A revised figure can be computed when all the reports have been obtained. If an industry is expanding and new firms are appearing, it is, of course, desirable to include them.   Increased employment and production may

result from increased activity of existing firms or the appearance of new ones. Similarly, firms may cease to exist and must be dropped from a reporting list.

Another source of incomparability may be that the unit of reporting has changed. If it is merely a question of changing from a pound basis to a ton basis, this is a simple matter. Where the product has changed in kind, however, it is difficult to find a satisfactory solution. How, for instance, can we compare the physical production of radios between 1925 and 1935? Not only was there a difference in the proportion of radios of different grades sold in the two years, but radios that were the same with respect to price, weight, number of tubes, or any other readily measurable characteristic, were still vastly different in their capacity to render utility to the consumer.

Selected References

E. C. Bratt· *Business Cycles and Forecasting*, Chapter IV; Business Publications, Inc, Chicago, 1937.

F. E. Croxton and D. J. Cowden· *Practical Business Statistics*, Chapter XIII; Prentice-Hall, Inc., New York, 1934.

E. E. Day: *Statistical Analysis*, Chapters XV and XIX; Macmillan Co., New York, 1927.

C. O. Hardy and G. O. Cox: *Forecasting Business Conditions*, Chapter I; Macmillan Co., New York, 1927.

W. C. Mitchell: *Business Cycles, the Problem and Its Setting;* National Bureau of Economic Research, New York, 1928. This book should be read in its entirety by anyone seriously interested in time series analysis. Chapter III is on statistical analysis, while Chapter V contains a summary of the entire book.

J. G. Smith: *Elementary Statistics*, Chapter XI; Henry Holt and Co., New York, 1934. A non-mathematical treatment of theories of time series.

Carl Snyder *Business Cycles and Business Measurements*, Chapter I; Macmillan Co., New York, 1927. Contains historical material, partly non-quantitative

CHAPTER XV

ANALYSIS OF TIME SERIES
SECULAR TREND

Objects and Method

There are two important reasons for attempting to describe the trend of a series by some kind of curve. First, it may be desired to measure the deviations from trend. These deviations consist of cyclical. seasonal, and accidental movements. Frequently the obtaining of these deviations is but one step in attempting to isolate cycles, in order to study them. Second, it may be desired to study the trend itself, in order to note the effect of factors bearing on the trend, to compare one trend with another, to discover what effect trend movements have on cyclical fluctuations, or to forecast future trend movements.

The purpose for which measurements are made partly determines the methods adopted. If the object is solely to isolate cycles, it seems reasonable to suppose that the trend line chosen should pass through the cycles in such a way as approximately to allow a balancing between the positive and negative phases of each cycle. Whether a curve is deemed to have accomplished this object depends, of course, upon our conception of what constitutes a cycle in each case. If, on the other hand, the object is to make comparisons, generalizations, or forecasts, the curve should be not only logical, but also of such a nature that it can readily be expressed by a mathematical formula. By so doing, a person can, for instance, say that at a given time a series shows a certain rate or a certain amount of growth per annum, and that, if this tendency continues, the trend will reach a certain value at some specified time in the future. Fitting a trend by a mathematical formula does not, however, remove the subjective element from trend fitting. The statistician can vary somewhat the shape of the curve by selection of the type of formula he employs, or the years to which he fits the curve. It remains true, therefore, that the statistician decides in advance, *upon as objective and logical a basis as possible*, what he thinks the trend ought to look like, and then selects the mathematical method that will closely approximate this result.

Trend Fitted by Inspection

The simplest method of describing a trend graphically is merely to draw it freehand, or perhaps to make use of a transparent ruler or a French curve. It is well to plot the data on semi-logarithmic paper also, for the trend may tend to straighten out if so plotted. The trend will be a straight line on this type of paper if the series is increasing or decreasing at a constant rate.

An attempt was made by one of the writers to fit, by inspection, a trend to the annual rayon consumption data: the results are shown in Chart 149.
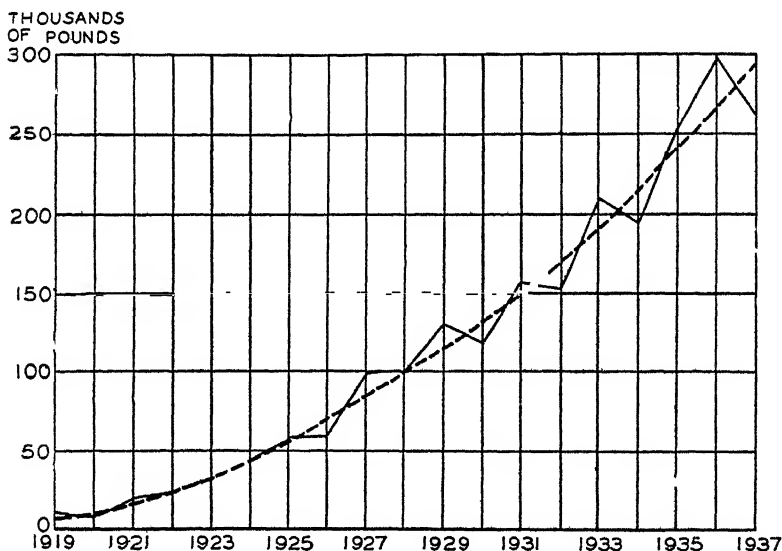
Chart 149.   Trend Line Fitted by Inspection to Total Domestic Consumption of Rayon in the United States, 1919-1937.   (Data of Table 80.)

This highly subjective method is open to the objection that may be made to all subjective methods—the statistician determines what answer he wants and then proceeds to obtain it. But, as has been said, he can accomplish very nearly the same result merely by careful selection from among numerous available mathematical methods. The most valid objection is that subjective methods require the exercise of a high level of judgment and the statistician therefore cannot safely have such work done by hired clerks.

Moving Averages

Simple moving averages and annual trend values. A simple and flexible mathematical method of trend fitting is the moving average. The process

of computing a 3-year moving average is shown in Table 80, United States rayon consumption figures being used for illustrative purposes. In column 4, the average 12,587 is $\dfrac{9,291 + 8,718 + 19,751}{3}$; 17,739 is the average of 8,718, 19,751, and 24,747; and so on. It should be noted that the moving average figures are placed opposite the center of the 3-year periods to which they refer, for the same reason that figures referring to a whole

TABLE 80

COMPUTATION OF 3-YEAR MOVING AVERAGE OF UNITED STATES CONSUMPTION OF RAYON, 1919–1937

(Thousands of pounds)

Year (1)	Consumption (2)	3-year moving total (3)	3-year moving average (4)
1919	9,291		.
1920	8,718	37,760	12,587
1921	19,751	53,216	17,739
1922	24,747	77,056	25,685
1923	32,558	99,548	33,183
1924	42,243	133,078	44,359
1925	58,277	161,150	53,717
1926	60,630	218,955	72,985
1927	100,048	260,779	86,926
1928	100,101	331,597	110,532
1929	131,448	349,517	116,506
1930	117,968	406,776	135,592
1931	157,360	427,369	142,456
1932	152,041	521,284	173,761
1933	211,883	558,695	186,232
1934	194,771	659,330	219,777
1935	252,676	745,041	248,347
1936	297,594	811,465	270,488
1937	261,195	.	.

Source: Textile Economic Bureau, *Rayon Organon*, Vol IX, No 2 January 21, 1938, p 16

period are customarily plotted on a chart in the middle of the appropriate spaces.

Chart 150 shows three different moving averages fitted to these data. All are bad fits. The 3-year moving average traces an inverse cycle. Because cycles in the rayon industry appear to last about two years, a 3-year moving average always includes either two depression years and one year of prosperity, or two good years and one bad year. On the other hand, the 5-year moving average, since it always includes either two cyclical peaks and three troughs, or three peaks and two troughs, dips down into

the troughs and reaches up into the peaks. The 7-year moving average follows the same general pattern as the 3-year, but is smoother.

From the reasoning of the preceding paragraph it must be obvious that, unless the moving average is the same length as the movement being smoothed (or some integral multiple thereof), the moving average will vary either directly or inversely with the undulations of that movement. If, however, it is an integral multiple of the wave length of the series—say, once, twice, or three times the average number of years or months in such movements—any particular moving average value will contain the same number of peaks as troughs, and the moving average will tend to smooth
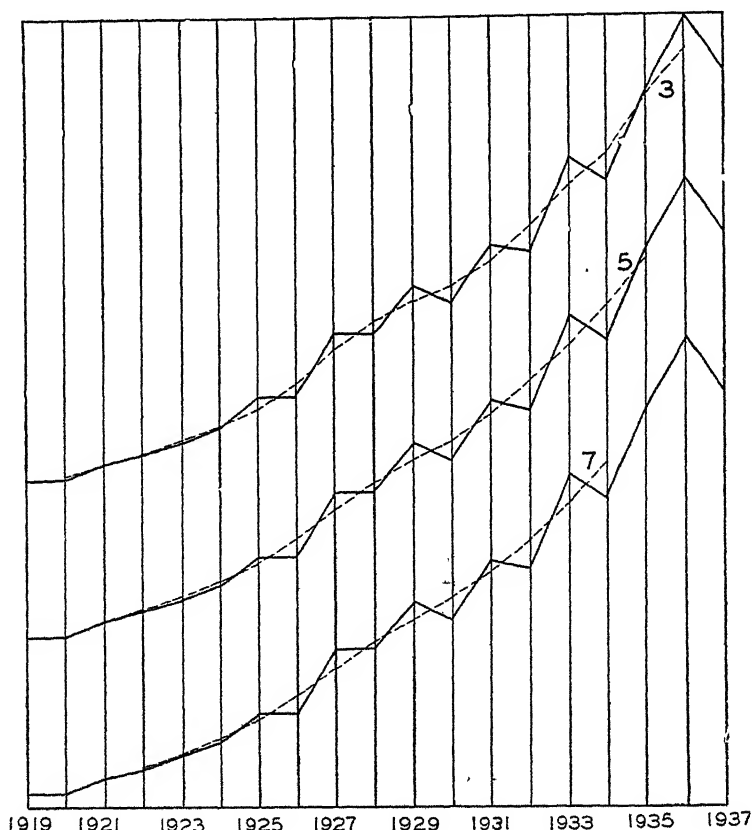
Chart 150. 3-Year, 5-Year, and 7-Year Moving Average Trends Fitted to Rayon Consumption, 1919-1937. (For purposes of comparison the different curves have been plotted close together on the same chart instead of on separate charts. Each curve is plotted to the same vertical scale, but at a different level. This arrangement is sometimes referred to as a multiple axis chart. For original data and 3-year moving average see Table 80.)

out the fluctuations which are sought to be eliminated. This principle, that *the period of the moving average should conform to the duration of the movement being smoothed,* is probably the most important one to observe in the use of moving averages.

Since the duration of rayon cycles appears to be about two years, it would therefore be better to try a moving average of two or four years. An extra step, however, is involved in this process, illustrated by Table 81. In column 3 of this table, 18,009, which is the total of the two years 1919 and 1920, is placed between these two years; and in like manner is 28,469 placed between 1920 and 1921. A 2-year moving average is then taken of the 2-year moving total in order to center the figures opposite, rather than between, years. The computation of a centered 4-year moving average is shown by Table 82. The procedure is similar to that in Table 81. except that two short cuts are used. First, as a matter of convenience, column 3 figures are placed opposite second years instead of between second and third years. Second, in order to eliminate one series of divisions, each consecutive pair of moving total figures is added together and the result is divided by 8 (or, to save work, multiplied by the reciprocal of 8, which is .125). This gives the moving average figures of column 5. It should be noticed that the first entry in this column is opposite 1921, which is the third year.

A glance at Chart 151 reveals that either the 2- or the 4-year moving average is superior to the 3- or 5-year moving average. The 2-year moving average passes through the center of each cycle. Mathematically this must be true since the cycles are two years in length. It may appear to the reader that what we describe as a centered 2-year moving average is really a 3-year moving average with the middle year given double weight. This is true, but it may also be regarded as an estimate of consumption during a 2-year period consisting of the middle year plus the half year preceding and the half year following. Although the 2-year moving average goes through the center of each cycle, it has more bends in it than are ordinarily considered appropriate for a trend. The 4-year moving average, though not passing so closely through the centers of cycles, is smoother, and on the whole perhaps better. On the other hand, the 6-year moving average is undesirable since it seems to lie above a reasonable trend for the years 1922, 1923, and 1924. It is inevitably true that a moving average, unless the items are elaborately weighted, will smooth out not only the undesirable irregularities, but also part of the curve which it is sought to approximate. Thus the moving average will fall below a trend which is concave downward and above one which is concave upward (as shown in the chart shown on page 393). Furthermore, the 6-year moving average eliminates (or necessitates estimates for) the first three and the last

three years, leaving only thirteen of the nineteen years with trend values!
Of course, the curve can be extended freehand in each direction, a highly

TABLE 81

Computation of Centered 2-Year Moving Average of United States Rayon
Consumption, 1919–1937

(Thousands of pounds)

Year	Consumption	2-year moving total	2-year moving average	2-year moving total of 2-year moving average	Centered 2-year moving average
(1)	(2)	(3)	(4)	(5)	(6)
1919	9,291			. . .	. . .
		18,009	9,004.5		
1920	8,718			23,239.0	11,620
		28,469	14,234.5		
1921	19,751			36,483.5	18,242
		44,498	22,249.0		
1922	24,747			50,901.5	25,451
		57,305	28,652.5		
1923	32,558			66,053.0	33,026
		74,801	37,400.5		
1924	42,243			87,660.5	43,830
		100,520	50,260.0		
1925	58,277			109,713 5	54,857
		118,907	59,453.5		
1926	60,630			139,792 5	69,896
		160,678	80,339.0		
1927	100,048			180,413.5	90,207
		200,149	100,074.5		
1928	100,101			215,849.0	107,924
		231,549	115,774.5		
1929	131,448			240,482 5	120,241
		249,416	124,708 0		
1930	117,968			262,372.0	131,186
		275,328	137,664 0		
1931	157,360			292,364.5	146,182
		309,401	154,700 5		
1932	152,041			336,662 5	168,331
		363,924	181,962 0		
1933	211,883			385,289.0	192,644
		406,654	203,327.0		
1934	194,771			427,050 5	213,525
		447,447	223,723.5		
1935	252,676			498,858.5	249,429
		550,270	275,135.0		
1936	297,594			554,529.5	277,265
		558,789	279,394.5		
1937	261,195			. . .	. . .

Source: See Table 80.

subjective procedure, or a remedy may be adopted that is nearly as bad as the disease itself. Estimates can be made of values for the original

TABLE 82

COMPUTATION OF CENTERED 4-YEAR MOVING AVERAGE OF UNITED STATES RAYON CONSUMPTION, 1919–1937

(Thousands of pounds)

Year (1)	Consumption (2)	4-year moving total (3)	2-year moving total of 4-year moving total (4)	Centered 4-year moving average [Col. 4 ÷ 8] (5)
1919	9,291		..	..
1920	8,718	62,507	...	...
1921	19,751	85,774	148,281	18,535
1922	24,747	119,299	205,073	25,634
1923	32,558	157,825	277,124	34,640
1924	42,243	193,708	351,533	43,942
1925	58,277	261,198	454,906	56,863
1926	60,630	319,056	580,254	72,532
1927	100,048	392,227	711,283	88,910
1928	100,101	449,565	841,792	105,224
1929	131,448	506,877	956,442	119,555
1930	117,968	558,817	1,065,694	133,212
1931	157,360	639,252	1,198,069	149,759
1932	152,041	716,055	1,355,307	169,413
1933	211,883	811,371	1,527,426	190,928
1934	194,771	956,924	1,768,295	221,037
1935	252,676	1,006,236	1,963,160	245,395
1936	297,594	...	...	...
1937	261,195	...	...	..

Source. See Table 80

data for several years on either side of the actually known data. In this case estimates may be made for 1916, 1917, 1918, and 1938, 1939, 1940. The 6-year moving average will then be affected by these hypothetical data, and the trend will run from 1919 through 1936. There is grave doubt concerning the validity of this procedure, but some authorities hold that, although we do not know with certainty concerning these periods, it is better to use what partial knowledge we possess than to use none whatever.

Obtaining monthly trend values from annual moving averages. In the above illustrations, moving averages have been fitted to annual data.
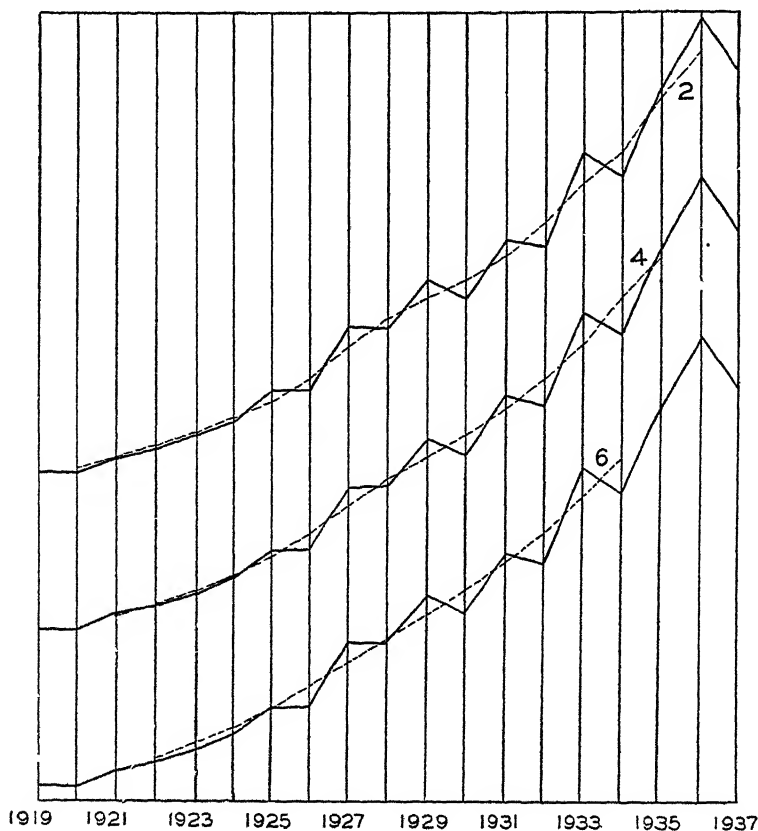
Chart 151. Centered 2-Year, 4-Year, and 6-Year Moving Average Trends Fitted to Rayon Consumption, 1919-1937. (For purposes of comparison the different curves have been plotted close together on the same chart instead of on separate charts. Each curve is plotted to the same vertical scale, but at a different level. This arrangement is sometimes referred to as a multiple axis chart. For original data and 2-year and 4-year moving averages see Tables 81 and 82.)

Monthly values can be obtained by interpolation between annual averages. For instance, in Table 83 the 1927 and 1928 trend values are 88,910 and 105,224 respectively. The increase is 16,314 during the intervening year, or 1,359.50 per month. Technically, 88,910 should be centered between June and July 1927, the July value being $88,910 + (1,359.50 \div 2) = 89,589.75$. The monthly trend values are then obtained by successive additions or subtractions of 1,359.50 (see Table 83).

If it is desired to fit a moving average trend to monthly data that have not been adjusted for seasonal variation, we should take some integral multiple of twelve months (approximating the average duration of the cycle); otherwise remnants of seasonal movements, either positive or inverse, will appear in the trend. For the rayon data, 24 or 48 months would be appropriate. Such a method gives fairly satisfactory
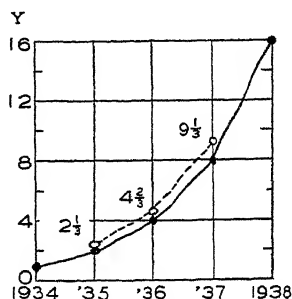
Three-Year Moving Average (Broken Line) of a Series of Data Which Are Concave Upward.

TABLE 83

INTERPOLATING ANNUAL MOVING AVERAGES TO OBTAIN MONTHLY TREND VALUES OF UNITED STATES RAYON CONSUMPTION, JULY 1927 THROUGH JUNE 1928

(Thousands of pounds)

Year and month	Annual trend values	Monthly trend values
1927·		
June　.. ...	..	.
	88,910	
July　.　..	..	89,589.75
August　.　　..	.	90,949 25
September　.	.	92,308 75
October　.　..	.	93,668.25
November .　. .	..	95,027 75
December.	..	96,387 25
1928.		
January　.	...	97,746 75
February　.　..	. .	99,106 25
March　..　...	..	100,465 75
April	..	101,825 25
May　...... .	..	103,184.75
June　.. . 　.	..	104,544 25
	105,224	
July	...	...

Source: Table 82.

results for these data, although, being an even number of periods, the moving average requires centering. If, however, the cycles averaged $3\frac{1}{2}$ years in length, it would be necessary to use $7 \times 12 = 84$ months in the moving average. More difficult would be the problem if the average cycle length were some inconvenient figure such as $3\frac{5}{8}$ years, which would require 276 months to fulfill strictly the requirements laid down above!

Of course, if the data have previously been *adjusted for seasonal movements* (see Chapter XVII), this difficulty disappears. Again taking the rayon data, turning points of which may be read from Chart 203, page 555, we have computed the average cycle length to be 24.45 months, as is shown by the following analysis. In order to avoid centering, a period of 25 months would therefore seem appropriate.

COMPUTATION OF AVERAGE CYCLE LENGTH OF RAYON
CONSUMPTION

Turning point	Peak to peak (Months)	Trough to trough (Months)
Peak: April 1923	..	.
Trough: February 1924.	..	..
Peak: May 1925	26	..
Trough: June 1926	...	28
Peak: May 1927	24	.
Trough: July 1928 . .	..	25
Peak: June 1929	25	.
Trough: October 1930.......	...	27
Peak: May 1931.	23	..
Trough: May 1932	...	19
Peak: June 1933	25	
Trough: September 1934......	.	28
Peak: July 1935....	25	..
Trough: March 1936	...	18
Average length........	24 7	24.2

Moving averages: summary. From what has been said it should be apparent that the fitting of a trend by a moving average is a procedure requiring the exercise of considerable judgment. It may be well therefore to conclude by summarizing some of the characteristics of moving averages.

(1) A moving average smoothes out fluctuations, provided its period is some integral multiple of the length of the movement to be smoothed. Accordingly, irregular movements are usually smoothed by a 3- or 5-month moving average, seasonals by a 12-month moving average, and cycles by a somewhat longer moving average. Since cycles usually vary in duration with the passage of time, this casts doubt on the appropriateness of a moving average trend for certain series.

(2) If the moving average is for an even number of periods, the resulting values must be centered by a 2-period moving average.

(3)   The larger the number of items used in the average, the smoother it becomes, for the less important, relatively, becomes any item which is added or dropped.

(4)   The larger the number of items, the greater the tendency to iron out, not only the fluctuations smaller in duration than the period of the average, but also part of the curvature of a non-linear trend itself.

(5)   The larger the number of items in the period, the greater the number of trend values on each end which must remain unknown, or estimated Specifically, if the average embraces N periods, there will be $\dfrac{N-1}{2}$ trend values omitted at each end; but $\dfrac{N}{2}$ if N is even and the moving average is centered.

(6)   A moving average is a descriptive measure rather than a summary measure.   Thus, to say that a trend is described by an eleven-year moving average throws no light on the way in which the series grows or declines. Also, a moving average states no "law" of change.   Nevertheless, the moving average is useful as a step preliminary to deciding on the final type of trend, and as a final step when the trend is not well defined or does not conform to any reasonably simple mathematical equation type.

Straight Line Trend

A mathematical equation not only is a descriptive measure of the trend, but also gives a concise definition of that trend.   If the trend itself is to be studied, or is to be extended beyond the data, it is especially desirable that it be so determined that it can be described by a mathematical equation.

Description.   The simplest type of curve is the straight line, which is described by an equation of the type $Y_C = a + bX$, in which X is the independent variable and Y_C the trend value of the dependent variable.[1] Since their values must be determined for each of the series being analyzed, a and b are referred to as *unknowns*. They are also called *constants* since, once their values are determined, they do not change.

To take the simplest case, suppose that $a = 0$ and $b = 1$.   The equation then becomes: $Y_C = X$; and this means that with each increase of one unit of the independent variable, the dependent variable also increases one unit.   This equation is plotted in the upper left-hand section of Chart 152.   Incidentally, it should be observed that all four quadrants are shown

---

[1] The symbol Y will be used to designate an observed value of the dependent variable; while Y_c indicates a value that has been computed, usually from a mathematical equation.
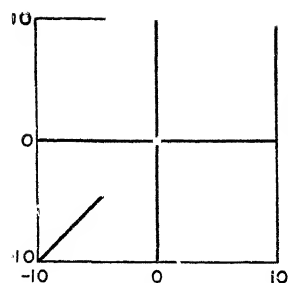
X	Y
-10	-10
- 5	- 5
0	0
5	5
10	10

Y = X
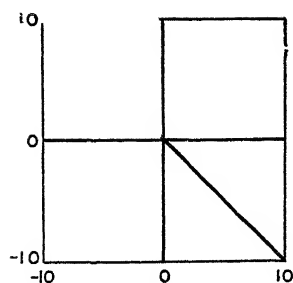
X	Y
-10	10
- 5	5
0	0
5	-5
10	-10

Y = -X
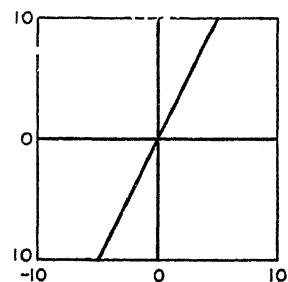
X	Y
-5	-10
-3	- 6
0	0
3	6
5	10

Y = 2X
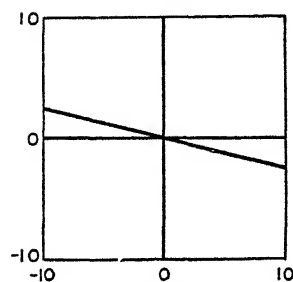
X	Y
-8	2
-4	1
0	0
4	-1
8	-2

Y = -.25X
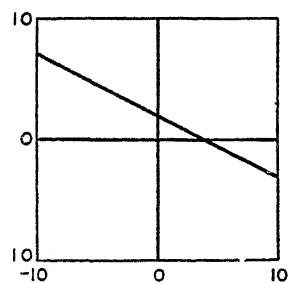
X	Y
-10	7
- 6	5
0	2
6	-1
10	-3

Y = 2 - .5X
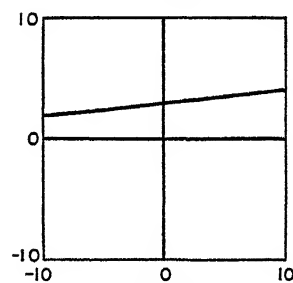
X	Y
-10	2
- 5	2.5
0	3
5	3.5
10	4

Y = 3 + .1X
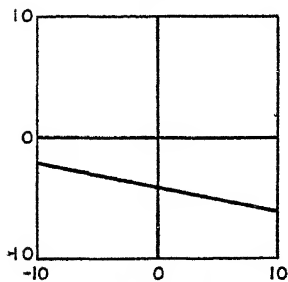
X	Y
-10	-2
- 5	-3
0	-4
5	-5
10	-6

Y = -4 -.2X
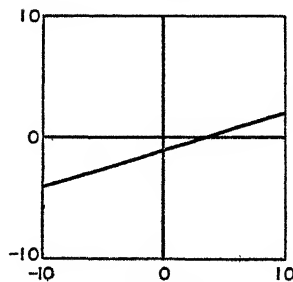
X	Y
-10	-4
- 5	-2.5
0	-1
5	5
10	2

Y = -1 +.3X

Chart 152   Straight-line equations and curves.

in this chart. Before attempting to plot a curve, it is well to draw up a table of X and Y_C values, as shown on the chart, in which are recorded the computed values of Y that correspond to selected values of X. As a matter of fact, only two points are needed to plot this or any straight line, and most accurate results are obtained by using two X values a considerable distance from each other.

Other straight line equations and their curves are shown in the other sections of Chart 152, an inspection of which yields the following information: a is the value of Y when X is 0 (the Y value at the X origin), or, as it is frequently termed, the Y intercept; while b indicates the steepness, or slope, of the line. When b is positive, the slope is upward; when b is negative, the slope is downward.

TABLE 84

COMPUTATION OF SEMI-AVERAGES TREND FOR ELECTRIC POWER
PRODUCTION, 1921–1930

(Millions of kilowatt hours)

Year (1)	Average monthly production (2)	Semi-averages (3)	Trend values (4)
1921	3,415	. . .	3,380 0
1922	3,971	.	3,933.2
1923	4,639	4,486.4 ✓	4,486.4
1924	4,918	.	5,039 6
1925	5,489	.	5,592 8
1926	6,149	. . .	6,146.0
1927	6,684	. .	6,699.2
1928	7,321	7,252 4 ✓	7,252.4
1929	8,113	. . .	7,805.6
1930	7,995	. .	8,358 8

Source· United States Department of Commerce, *Survey of Current Business,*
1936 Supplement, p 85

Method of selected points. Since the location of only two points is necessary to obtain a straight line equation, it is obvious that we may select two representative points and connect them by a straight line. Of course, this method is highly subjective. Typically, the data are divided into halves, and an average is computed for each half. This is known as the method of *semi-averages*, and is illustrated in Table 84, dealing with electric power production. As indicated by column 3, the trend value for 1923 is 4,486.4. and that for 1928 is 7,252.4. This is an increase during

the 5-year period of 7,252.4 − 4,486.4 = 2,766.0.   The annual increment is of course one-fifth of that amount, or 553.2.   It is therefore apparent that this trend can be described mathematically by the equation $Y_C = 4,486.4 + 553.2X$, with origin at 1923.   The trend values are most easily found by successively adding 553.2 to 4,486.4, except that, for 1922 and 1921, subtractions are involved instead.   As indicated by Table 84, the 1921 trend value is 3,380.0.   The equation could therefore be written $Y_C = 3,380.0 + 553.2X$, with origin at 1921.
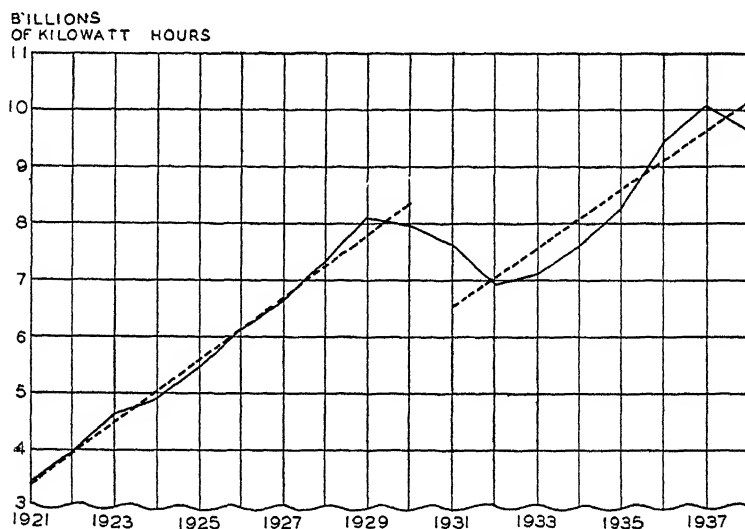
Chart 153.   Straight-Line Trends Fitted by Method of Semi-Averages to Average Monthly Production of Electric Power, 1921-1930 and 1931-1938.   (Original data from U S. Department of the Interior, *Geological Survey,* as quoted by U. S. Department of Commerce, *Survey of Current Business:* 1936 Supplement, p. 85; Vol. 17, February 1937, p. 41; Vol. 18, March 1938, p. 81.)

This method is to be commended for its simplicity and is used to some extent in practical work, but nearly all statisticians prefer the more refined method described in the ensuing pages.[2]
The results of the semi-average procedure are shown in Chart 153.   This chart shows two trends fitted by the method of semi-averages: one is fitted to 1921–1930 data and the other to 1931–1938 data.   This sharp break in the trend may seem to the reader to be inconsistent with the idea of what a trend really is.   Many economists believe that continuity and sta-

---

[2] The method referred to is the method of least squares: For the same data it gives the equation $Y_C = 3,426\ 25 + 542\ 922X$, with origin at 1921, which compares with $Y_c = 3,380.0 + 553.2X$ by the method of semi-averages.

bility constitute the very essence of the trend.   On the other hand, there are some who believe that the depression which reached its trough in 1932–1933 was not a mere business cycle.   It represented a breakdown in our economic order.   The old trends did indeed continue, but from a lower level; or we can say that they were set back about five years.   Chart 153 illustrates this concept, although it does not necessarily commit the authors to this view.

Method of least squares.   A more refined method, which can be applied to more complex types of trends, is favored by most statisticians.   This method is designed to accomplish two results.
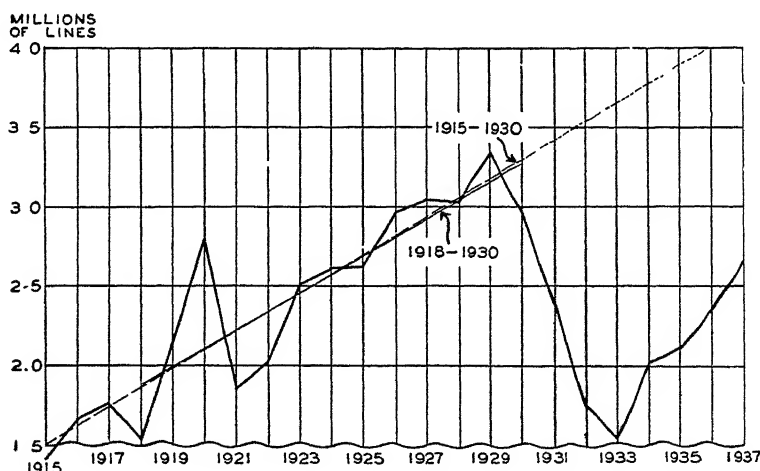
Chart 154.   Straight-Line Trends Fitted by Method of Least Squares to United States Magazine Advertising.   (Data of Tables 88 and 89.)

1.   *The sum of the vertical deviations from the straight line must equal zero.* If we should connect by a vertical line each production figure (1915–1930) in Chart 154 with the dashed trend, the vertical lines extending upward from the trend would exactly balance those extending downward. It would not suffice, however, merely to set the sum of the deviations equal to zero, since any straight line (other than vertical) passing through $\overline{X}$, $\overline{Y}$ would fulfill the requirement.   This condition would even be satisfied by a straight line sloping at right angles to the trend.

2.   *The sum of the squares of all these deviations, both above and below the trend line, must be less than the sum of the squares from any other conceivable straight line.*   It is because of this second characteristic of such a line that the method of fitting it to obtain this result is called the method of

least squares.[3]   In fitting a curve to meet the second requirement, the first requirement is automatically satisfied.[4]

In a sense, then, a trend line fitted by the method of least squares is analogous to the arithmetic mean, for the latter measure is a *single value*, rather than a *series of values*, summarizing a statistical series which possesses the two characteristics mentioned above.

TABLE 85

(Magazine advertising in thousands of lines per month)

Year	X	Advertising Y	Observation equation $Y = a + bX$
1918	0	1,547	$1,547 = a$
1919	1	2,142	$2,142 = a + b$
1920	2	2,803	$2,803 = a + 2b$
1921	3	1,856	$1,856 = a + 3b$
1922	4	2,030	$2,030 = a + 4b$
1923	5	2,520	$2,520 = a + 5b$
1924	6	2,620	$2,620 = a + 6b$
1925	7	2,623	$2,623 = a + 7b$
1926	8	2,958	$2,958 = a + 8b$
1927	9	3,038	$3,038 = a + 9b$
1928	10	3,032	$3,032 = a + 10b$
1929	11	3,384	$3,384 = a + 11b$
1930	12	2,984	$2,984 = a + 12b$

Source United States Department of Commerce, *Survey of Current Business*, October 1933, p 20, 1936 Supplement, p 24, December 1936, p 25, May 1937, p 26

---

[3] A distribution of chance errors follows the normal curve; and it can be demonstrated that the greatest probability of obtaining deviations from some computed value or series of values which are distributed in this fashion is obtained when the sum of the squared deviations is at a minimum (see Appendix B, section XV–2). If the deviations are distributed in this fashion, then the fitted value is most probable. If it is believed that deviations from the appropriate norm are chance errors, it follows that the method of least squares is the appropriate method of fitting The method is also convenient algebraically, as the student can observe in connection with correlation analysis and analysis of variance Time series fluctuations around a trend line are not, however, independent accidental occurrences, and it is to be doubted that there is any special reason for using the method of least squares in trend fitting, other than for its convenience. Certain of the trends explained in this volume are, in fact, fitted by other methods Some statisticians even argue that the least squares criterion is not appropriate for time series trends, since time series are sometimes characterized by extreme deviations not in accordance with the normal law The method of least squares, of course, is particularly influenced by extreme deviations because of the squaring process

[4] The mean of the Y_c values is the same as the mean of the Y values. This is demonstrated in Appendix B, Section XXII—1 Before reading that explanation, however the reader should peruse the remainder of this section.

Normal equations. In order to fit a straight line by the method of least squares, two *normal equations* must be obtained and solved simultaneously, since there are two constants or unknowns to be found. The normal equations are obtained from *observation equations*. There are as many observation equations as there are observations. In this case we shall use United States magazine advertising 1918–1930 as an illustration; hence there are thirteen observations. Since the straight line trend is of the equation type $Y_C = a + bX$, we can insert the observed values of X and of Y in the expression $Y = a + bX$ and obtain the thirteen observation equations, as in Table 85. In this table 1918 has been taken as the X origin, although any other year could have been set down as zero.

Each original observation equation is now multiplied by the coefficient of a. Since the coefficient of a is 1 for each equation, the resulting observation equations, which are shown in column 3 of Table 86, are unchanged.

TABLE 86

DERIVATION OF NORMAL EQUATIONS FROM OBSERVATION EQUATIONS FOR UNITED STATES MAGAZINE ADVERTISING DATA

Original observation equations $Y = a + bX$ (1)	Coefficient of a (2)	First set of observation equations $Y = a + bX$ [Col. 1 × Col. 2] (3)	Coefficient of b X (4)	Second set of observation equations $XY = aX + bX^2$ [Col. 1 × Col 4] (5)
$1,547 = a$	1	$1,547 = a$	0	
$2,142 = a + b$	1	$2,142 = a + b$	1	$2,142 = a + b$
$2,803 = a + 2b$	1	$2,803 = a + 2b$	2	$5,606 = 2a + 4b$
$1,856 = a + 3b$	1	$1,856 = a + 3b$	3	$5,568 = 3a + 9b$
$2,030 = a + 4b$	1	$2,030 = a + 4b$	4	$8,120 = 4a + 16b$
$2,520 = a + 5b$	1	$2,520 = a + 5b$	5	$12,600 = 5a + 25b$
$2,620 = a + 6b$	1	$2,620 = a + 6b$	6	$15,720 = 6a + 36b$
$2,623 = a + 7b$	1	$2,623 = a + 7b$	7	$18,361 = 7a + 49b$
$2,958 = a + 8b$	1	$2,958 = a + 8b$	8	$23,664 = 8a + 64b$
$3,038 = a + 9b$	1	$3,038 = a + 9b$	9	$27,342 = 9a + 81b$
$3,032 = a + 10b$	1	$3,032 = a + 10b$	10	$30,320 = 10a + 100b$
$3,384 = a + 11b$	1	$3,384 = a + 11b$	11	$37,224 = 11a + 121b$
$2,984 = a + 12b$	1	$2,984 = a + 12b$	12	$35,808 = 12a + 144b$
Normal equation		$33,537 = 13a + 78b$		$222,475 = 78a + 650b$

Source. Table 85

Next, each original observation equation is multiplied by the coefficient of b, with the results shown in column 5. We now have two new sets of observation equations. Each of these sets is summed to obtain the normal equations shown at the bottom of the table. The two normal equations

$$\text{I.} \quad 33,537 = 13a + 78b,$$
$$\text{II.} \quad 222,475 = 78a + 650b,$$

are the numerical representation of the following equation types:[5]

I. $\Sigma Y = Na + b\Sigma X,$
II. $\Sigma XY = a\Sigma X + b\Sigma X^2.$

In order to solve these two equations simultaneously, we may multiply equation I by 6 ($= 78 \div 13$) and subtract equation I from equation II, thus obtaining an equation with one unknown, b:

I. $201,222 = 78a + 468b$
II. $222,475 = 78a + 650b$

$21,253 = \qquad\qquad 182b$
$b = 116.7747$

Having obtained b, we obtain a by substituting the value of b in equation I. Thus

I. $33,537 = 13a + 78\,(116.7747)$
$a = 1,879.121.$

It is desirable to check the accuracy of the solution of the normal equations, either by obtaining a by substitution of b in normal equation II, or the values of a and of b may be substituted in the second normal equation as follows:

II. $222,475 = 78\,(1,879.121) + 650\,(116.7747)$
$= 222,474.99.$

The trend equation may now be written

$$Y_c = 1,879.121 + 116.774X,$$

with origin at 1918 and X units of 1 year.

Before proceeding with the discussion, let us summarize the general procedure for obtaining a trend equation of this type. (The procedure can be expanded for equations of higher degree.)

(1) Set up an equation for each observation by inserting the observed values of X and Y in the expression $Y = a + bX$.

(2) Multiply each original observation equation by the coefficient of the first unknown, a.

(3) Multiply each original observation equation by the coefficient of the second unknown, b.

(4) Sum each of the two resulting sets of observation equations. This gives the two normal equations.

(5) Solve the two normal equations simultaneously for b.

(6) Substitute the value of b in equation I and obtain the value of a

(7) Check the solution by substituting the values of a and b in equation II.

---

[5] For derivation of these normal equations, see Appendix B, section XV–1.

It is not necessary to set up an elaborate table such as Table 86. Table 87 is much simpler. ΣX need not be computed, nor is it necessary

TABLE 87

Straight Line Trend Fitted to Data of United States Magazine Advertising, 1918–1930

(Thousands of lines per month)

Year	X	Y	XY	Y_c
1918	0	1,547	0	1,879
1919	1	2,142	2,142	1,996
1920	2	2,803	5,606	2,113
1921	3	1,856	5,568	2,229
1922	4	2,030	8,120	2,346
1923	5	2,520	12,600	2,463
1924	6	2,620	15,720	2,580
1925	7	2,623	18,361	2,697
1926	8	2,958	23,664	2,813
1927	9	3,038	27,342	2,930
1928	10	3,032	30,320	3,047
1929	11	3,384	37,224	3,164
1930	12	2,984	35,808	3,280
Total	78	33,537	222,475	

Source: Table 86

Normal equations:

I. $\Sigma Y = Na + b\Sigma X$;   I. $33,537 = 13a + 78b$;

II. $\Sigma XY = a\Sigma X + b\Sigma X^2$ II. $222,475 = 78a + 650b$.

Trend equation:

$$Y_c = 1,879\ 121 + 116\ 7747X.$$

Origin, 1918. X units, 1 year.

to show an X^2 column in this table since ΣX and ΣX^2 can be looked up directly in Appendix M. In this appendix the sum of the first 12 natural numbers is shown to be 78, and the sum of the squares to be 650.[6] Table

---

[6] If we solve the two normal equations in symbolic form, a and b may be obtained as follows:

$$b = \frac{\Sigma XY - \overline{X}\Sigma Y}{\Sigma X^2 - \overline{X}\Sigma X};$$

$$a = \overline{Y} - b\overline{X}.$$

In the present instance

$$\overline{X} = \frac{78}{13} = 6.$$

$$\overline{Y} = \frac{33,537}{13} = 2,579.7692.$$

$$b = \frac{\Sigma XY - \overline{X}\Sigma Y}{\Sigma X^2 - \overline{X}\Sigma X} = \frac{222,475 - 6(33,537)}{650 - 6(78)} = \frac{21,253}{182} = 116\ 7747.$$

$$a = \overline{Y} - b\overline{X} = 2,579\ 7692 - (116.7747)6 = 1,879.121.$$

87 also provides an extra column for recording the trend values; they are obtained by the usual method of adding to a (1,879.121), successive amounts of b (116.7747) for each year after 1918.

Thus far, in the tables illustrating the method of least squares, we have taken the first year (1918) as the X origin. There is, however, an arithmetic advantage in having the origin at the middle year, which is the mean of the X values. It will be remembered that the sum of the deviations from the mean is zero. More generally, it may be stated, the terms of the normal equations containing sums of the odd powers of X become zero, an advantage that is especially important for curves of higher degree Thus for a straight line

I. $\Sigma Y = Na + b\Sigma X$ becomes $\Sigma Y = Na$;
II. $\Sigma XY = a\Sigma X + b\Sigma X^2$ becomes $\Sigma XY = b\Sigma X^2$.

The normal equations may also be written

$$\text{I.} \quad a = \frac{\Sigma Y}{N};$$

$$\text{II.} \quad b = \frac{\Sigma XY}{\Sigma X^2}.$$

The first normal equation in this form, therefore, merely states that a straight line, fitted by the method of least squares, passes through the point $\bar{X}$, $\bar{Y}$. The advantage in having the X values cancel out is that the two normal equations do not need to be solved simultaneously but can now be solved separately by a very simple process, thereby saving considerable labor.

Odd number of items. Table 88 illustrates the procedure of straight line trend fitting when the X origin is taken at the middle year and when an odd number of years is used. Zero is placed in the X column opposite 1924, and ΣX becomes zero. Again ΣX^2 is obtained from Appendix M. In this appendix the sum of the squares of the first six natural numbers is shown to be 91. Since we have six years on each side of 1924, the value of ΣX^2 is $2 \times 91 = 182$. As indicated below Table 88, the trend equation is

$$Y_C = 2{,}579.77 + 116.7747X,$$

with origin at 1924 and X units of one year. Observe that, after the trend equation is stated, two essential qualifying statements are recorded: (1) that the origin was taken at 1924, and thus the value of $X = 0$ in 1924, and $Y_C = a$ for that year; (2) that b has reference to the normal increment in production during one year. Were these statements not made, a person seeing the equation out of context might erroneously conclude that the normal production for 1918 (the first year of the series)

was 2,579.77 and that the normal *monthly* increase was 116.7747.   Table 88 again provides an extra column for recording the trend values.   They are obtained by adding to or subtracting from a (2,579.77) successive amounts of b (116.7747).   Note that, for 1918, $Y_C = 1,879$.   Obviously, therefore, if the origin be shifted to 1918, we may state the equation

$$Y_C = 1,879 + 116.7747X,$$

which is the same equation that was obtained previously.

TABLE 88

FITTING STRAIGHT LINE TREND: ODD NUMBER OF ITEMS

(United States magazine advertising data, thousands of lines per month)

Year	X	Y	XY		Y_c
1918	−6	1,547	− 9,282		1,879
1919	−5	2,142	−10,710		1,996
1920	−4	2,803	−11,212		2,113
1921	−3	1,856	− 5,568		2,229
1922	−2	2,030	− 4,060		2,346
1923	−1	2,520	− 2,520	−43,352	2,463
1924	0	2,620	0		2,580
1925	1	2,623	2,623		2,697
1926	2	2,958	5,916		2,813
1927	3	3,038	9,114		2,930
1928	4	3,032	12,128		3,047
1929	5	3,384	16,920		3,164
1930	6	2,984	17,904	64,605	3,280
Total	.	33,537	21,253		...

Source: Table 85

Normal equations:

$$\text{I.} \quad a = \frac{\Sigma Y}{N} = \frac{33,537}{13} = 2,579.77.$$

$$\text{II.} \quad b = \frac{\Sigma XY}{\Sigma X^2} = \frac{21,253}{182} = 116.7747.$$

Trend equation:

$$Y_c = 2,579.77 + 116.7747X.$$
Origin, 1924. X units, 1 year.

Even number of items.   The reader has probably already wondered whether the procedure described could be applied if the trend were to be fitted to a period with an even number of years, say 1915–1930.   The procedure is only slightly modified.   If the years 1915–1930 inclusive are used, the middle of the period falls between 1922 and 1923.   From this point of time it is one-half year to the middle of 1922 and one-half year to

the middle of 1923. Since it would be inconvenient to use fractions in the computations, however, one unit of the independent variable X is taken to represent six months. Therefore 1922 is labeled -1 and 1923 is labeled 1, as in Table 89. There is, of course, an interval of two 6-month periods

TABLE 89

COMPUTATION OF STRAIGHT LINE TREND: EVEN NUMBER OF ITEMS

(United States magazine advertising data, thousands of lines per month)

Year	X	Y	XY		Y_c
1915	−15	1,407	−21,105		1,508
1916	−13	1,669	−21,697		1,626
1917	−11	1,772	−19,492		1,745
1918	− 9	1,547	−13,923		1,864
1919	− 7	2,142	−14,994		1,983
1920	− 5	2,803	−14,015		2,102
1921	− 3	1,856	− 5,568		2,221
1922	− 1	2,030	− 2,030	−112,824	2,340
1923	1	2,520	2,520		2,458
1924	3	2,620	7,860		2,577
1925	5	2,623	13,115		2,696
1926	7	2,958	20,706		2,815
1927	9	3,038	27,342		2,934
1928	11	3,032	33,352		3,053
1929	13	3,384	43,992		3,172
1930	15	2,984	44,760	193,647	3,290
1931*	17	2,409	...		3,409
1932*	19	1,763	...		3,528
1933*	21	1,555	..		3,647
1934*	23	2,027	.		3,766
1935*	25	2,115	...		3,885
1936*	27	2,378	..		4,004
1937*	29	. .	.		4,122
Total		38,385	80,823		.

* X and Y values for years after 1930 are not used in computing trend
Source: See Table 85

Normal equations:

$$\text{I.} \quad a = \frac{\Sigma Y}{N} = \frac{38,385}{16} = 2399\ 06.$$

$$\text{II.} \quad b = \frac{\Sigma XY}{\Sigma X^2} = \frac{80,823}{1,360} = 59.4287.$$

Trend equation:

$$Y_c = 2,399\ 06 + 59.4287X.$$

Origin, 1922–1923. X units, $\frac{1}{2}$ year.

between any two points a year apart; therefore, 1921 is shown as −3, 1924 as 3, and so on. In obtaining a value for ΣX^2, in this case we must turn to Appendix N, which shows the sums of squares of odd natural

numbers. The sum of the squares of the first eight odd natural numbers (1, 3, 5, 7, 9, 11, 13, 15) is shown to be 680; ΣX^2 is twice that amount, or 1,360.

The trend equation as shown below Table 89 is

$$Y_c = 2{,}399.06 + 59.4287X,$$

with origin between 1922 and 1923, and X units of $\frac{1}{2}$ year. In obtaining the trend value for 1922, we must subtract 59.4287 from 2,399.06; but to obtain 1921, 1920, etc., we must successively subtract twice this amount, or 118.8574.

Since the main time series illustration running through this book will deal with magazine advertising for the period 1921 to date, a more convenient statement of the trend is with origin at 1921. The trend value for that year is shown by Table 89 to be 2,221 (to six digits, it is 2,220.77). We may therefore write the equation

$$Y_c = 2{,}220.77 + 118.8574X$$

with origin at 1921 and X units of one year. This permits us to obtain the trend values following 1921 by successive addition only. This is most easily done by putting 2,220.77 in the calculating machine and 118.8574 on the keyboard, and recording the trend value each time the value of b is added. The use of an adding machine necessitates inserting the value of b and subtotaling to obtain each trend value, but provides a record for checking against possible errors. Trend values from 1931 through 1937 are recorded in Table 89, although the data for these years were not used in obtaining the trend equation. Extending the trend in this fashion is a customary procedure, since it is not practical or desirable to recompute a complete new trend each year. Extension for 7 years on the basis of 16 years' experience is, of course, somewhat hazardous, but possibly the results are not so unreasonable as would have been obtained if the very unusual years of the great depression had been included in our computations. The fitted trend, with the extension, is shown in Chart 154. The portion of the trend that has been extended is dotted to distinguish it from the dashed line representing the trend values based upon observations. For purposes of comparison, the trend line fitted to the period 1918–1930 is also shown.

We now have two least-squares trends, one for the period 1918–1930, and another for the period 1915–1930. The question naturally arises: Which is better? Inspection of Chart 154 reveals that the difference between the two trends is so slight that for practical purposes it may be neglected. On logical grounds, however, the one fitted to the longer period is to be preferred, since measures become more reliable as the number

of observations is increased. Before leaving this section, it is well to lay down a few generalizations on this point.

(1) In order that the trend equation may be as reliable as possible, the maximum number of observations available should be used, provided there has been no change in the nature of the trend. If the nature of the trend has changed, a second trend (perhaps of a different type) should be fitted and perhaps spliced to the old one. The following two considerations modify this generalization somewhat.

(2) In order that the slope of the trend may be correct, it is important that the beginning and end of the series should not be at markedly different cyclical levels. If the first year is one of prosperity and the latter one of depression, the trend slope will have a downward bias. If the period begins with depression and ends with prosperity, the bias will be upward. Under any circumstances a period of extreme prosperity or depression near either end of the series will impart a bias to the slope in one direction or the other.

(3) In order that the trend as a whole may have the correct level, the period to which the trend is fitted should contain about the same area of prosperity as of depression. For instance, if the first and last years are peaks of extreme prosperity, the general level of the trend will be too high; if the first and last years are troughs of extreme depression, the general level will be too low.

The second and third considerations recede in importance as the period is increased in length. In the present instance it seemed desirable to have the first and last years ones of moderate recession. By so doing, a trend was obtained that is reasonable both as to general level and as to slope; that is, both a and b are reasonable.

Adaptation of equations to monthly data. Annual data have been used in fitting the straight line, although it may be desired to analyze monthly data. The process of fitting a trend to monthly data is not different from that of fitting it to annual data, but there are 12 times as many values to fit to, and the labor is multiplied by more than 12. It is therefore advisable to fit the trend to annual data, and then to transform it to a monthly basis. The difference between the two methods is usually negligible. In fact, it is probably better to use annual data than monthly data that has not been deseasonalized, since the presence of seasonal movement, if very violent, may distort the trend.

It will be recalled that the trend for the annual data for the years 1915–1930, with origin at 1921, is $Y_C = 2{,}220.77 + 118.8574X$. If the annual increment is 118.8574, the monthly increment is $118.8574 \div 12 = 9.90478$. However, there is an added difficulty: 2,220.77 is the value of

Y_C at the *end* of June 1921, whereas the monthly advertising data were considered as of the *middle* of the month. It is $5\frac{1}{2}$ months from the end of June 1921 to the middle of January 1921. The Y_C value for January 1921 therefore is $2,220.77 - 5.5(9.90478) = 2,166.29$, and the monthly equation with origin at January 1921 becomes $Y_C = 2,166.29 + 9.90478X$. Trend values for each month from January 1921 to December 1930 may now be obtained simply by adding successive increments of 9.90478 to 2,166.29. As a check, the December 1937 figures should be $2,220.77 + 197.5(9.90478) = 4,176.96$.

To summarize, we may say that, when the trend is fitted to annual averages of monthly data and when the number of years is odd, in order to convert the X units from years to months it is necessary to divide only the constant b by 12. If the trend has been fitted to an even number of

TABLE 90

UNITED STATES MAGAZINE ADVERTISING, EXPRESSED AS AN-
NUAL TOTALS AND AS AVERAGES PER MONTH,
BY YEARS, 1918–1930

Year	Total annual advertising (thousands of lines)	Average monthly advertising (thousands of lines)
1918	18,569	1,547
1919	25,702	2,142
1920	33,638	2,803
1921	22,271	1,856
1922	24,365	2,030
1923	30,233	2,520
1924	31,442	2,620
1925	31,473	2,623
1926	35,491	2,958
1927	36,453	3,038
1928	36,379	3,032
1929	40,606	3,384
1930	35,804	2,984

Source: See Table 85.

years, the value for b can be multiplied by 2 to convert it into change per year, after which the procedure is as above; or the transformation can be made directly, by dividing b by 6. In the foregoing illustration the trend has been fitted to *average monthly* values. Suppose, however, that the data were *total* magazine advertising for *each year*. A simple but somewhat laborious procedure is to divide the original data by 12 in order to reduce them to average monthly production, and then to proceed as above. If it is desired to fit the trend to the annual totals, a somewhat different procedure must be followed.

The data for United States magazine advertising can be set forth in either of two ways shown in Table 90: (1) as annual totals; (2) as monthly averages for each year.

Straight line equations with origin at 1924 are as follows:

Total annual advertising: $Y_C = 30,957.24 + 1,401.2964X$;
Average monthly advertising: $Y_C = 2,579.77 + 116.7747X$.

The second equation is the same as that shown below Table 88. It will be noticed that the values of a and b are exactly 12 times as great in the first equation as in the second. This is, of course, logical, since the first equation deals with annual totals, while the second deals with monthly averages. Thus, dividing the first equation by 12 changes a from the normal advertising for the *entire year*, 1924, to the normal advertising *per month* of 1924; and changes b from the normal annual increase in *total advertising per year* to the normal increase during the course of a year in *average monthly advertising*, as, for instance, the typical change in monthly advertising from *one January to the next January*. However, what is wanted is the typical change in monthly advertising from *one month to the next*. In other words, we have converted the Y units of measurement from an annual to a monthly basis, but not the X units. It is therefore necessary to divide b through again by 12 (or by 144 altogether). The equation now becomes

$$Y_C = 2,579.77 + 9.7312X,$$

with origin at the middle of 1924 (June 30) and X units of one month. To shift the origin to January 1921, which is 41.5 months before the middle of 1924, we compute

$$Y_C = 2,579.77 - 41.5 (9.7312) = 2,175.93,$$

which is the new a, and the equation in the desired form becomes

$$Y_C = 2,175.93 + 9.7312X,$$

with origin at January 1921 and X units of one month. To summarize, we may say that, when the trend is fitted to annual totals rather than averages of monthly data and when the number of years is odd, in order to convert the X and Y units from annual to monthly terms, it is necessary to divide the constant a by 12 and the constant b by 144. If the trend has been fitted to annual totals with an even number of years, the X units, as has been explained, are six months. The value for b can be multiplied by 2 to convert into changes per year, after which the procedure is as above. Or the transformation can be made directly, by dividing a by 12 and b by 72.

The tabular summary below is in convenient form for ready reference

when we wish to derive, from an equation fitted to annual data, an equation for use with monthly data.

Number of years	Type of data			
	Monthly averages		Annual totals	
	a	*b*	*a*	*b*
Odd	No change	Divide by 12	Divide by 12	Divide by 144
Even	No change	Divide by 6	Divide by 12	Divide by 72

Under all circumstances, shifting the origin to some convenient month involves an adjustment of half a month.

Other Simple Types of Trends

Series of curves. Occasionally no single curve will seem adequately to describe the trend. The electric power production series illustrated in Chart 153 is perhaps an illustration, but there the trend is shown as being discontinuous. A better illustration is Chart 155, in which are shown two connected straight lines fitted to average tractive power of steam locomotives. The first line is a least-squares fit to the 1923–1929 data, while the second is to the 1929–1935 data. This gives two trend values for 1929: 44.73 by the former equation, and 44.63 by the latter. In this case, inspection of the chart seems to show that a better trend would result if the latter figure were used. The splicing together of two trends is always a highly subjective procedure, and no general rules can be laid down for its accomplishment. The use of a series of curves is applicable not only to straight line trends but to any other type (or types) of curve. See, for instance, Chart 207. It is, however, better to avoid this method unless it is strongly supported by the appearance of the chart, and preferably also by the logic of the situation.

Related series as trend. Sometimes the trend of a series that has considerable amplitude of fluctuation can be described by the actual values of some related series that is more stable. Thus, bond yields are found to provide a not unreasonable trend for commercial paper rates, as shown in Chart 156. This method by itself, however, does not have very wide applicability. It should, of course, be used only when logically justified

A modification of this method, which is perhaps more widely applicable

is to utilize some other series as part of the trend. Thus, Warren M. Persons used *population growth* as one element in the trend of Barron's Annual Index of Production and Trade. His method is illustrated in Table 91. First the unadjusted index numbers are divided by figures representing the population of the United States relative to 1923–1925,
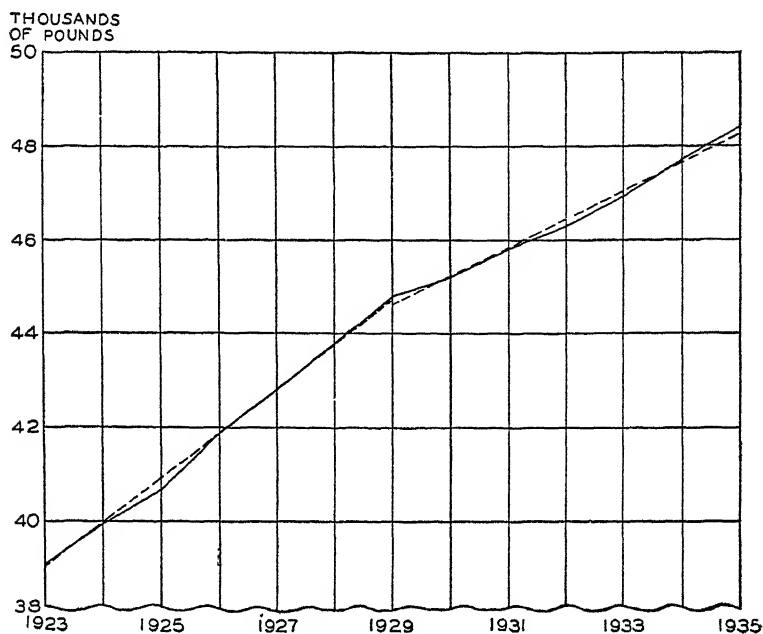
Chart 155. Series of Straight Lines as Trend of Average Traction Power of Steam Locomotives, 1923-1935. (Data from Committee on Public Relations of Eastern Railroads, *A Yearbook of Railroad Information, 1936 Edition*, p. 6.)

in order to obtain an index adjusted for population growth. This procedure has been referred to also in Chapters VII and XIV. The unadjusted index and the population curve are shown in part A of Chart 157, while the adjusted data and straight line trend are shown in part B of this chart. The straight line now seems to provide a good trend for these data. The straight line trend values are recorded in column 5 of Table 91. The trend values are the product of the two separate elements; that is, they are obtained by multiplying together the straight line values and the population relatives. The final results are shown in part C of Chart 157. The trend line, while not quite so smooth as the usual mathematical curve, seems to be a good description of the trend of this series.

Cyclical averages. When it is desired to measure cyclical deviations from a trend so obtained that the positive and negative portions of each
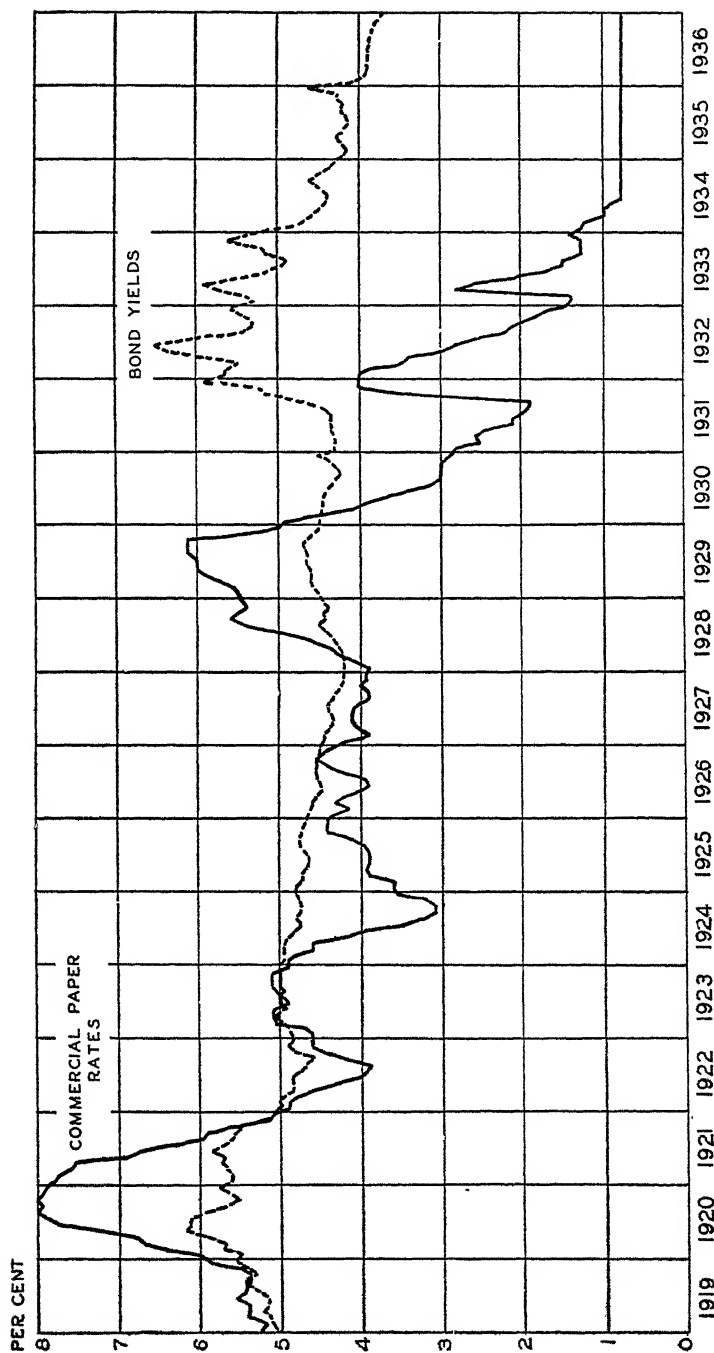
Chart 156. Bond Yields and Commercial Paper Rates, 1919-1936. (Data from Frederick R. Macaulay, *Bond Yields, Interest Rates, and Stock Prices*, Publication No. 33, National Bureau of Economic Research, New York, 1938, pp. A157–A161.)

413

POPULATION IN MILLIONS
INDEX IN PERCENTAGES
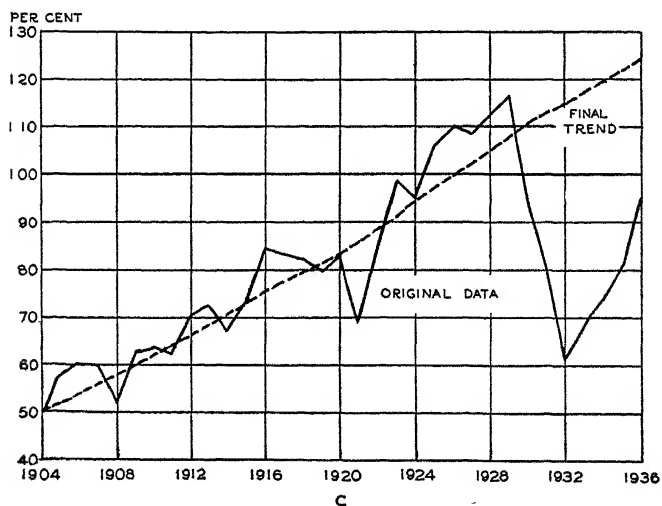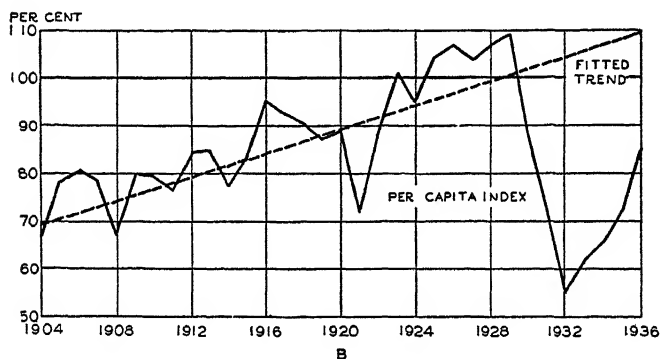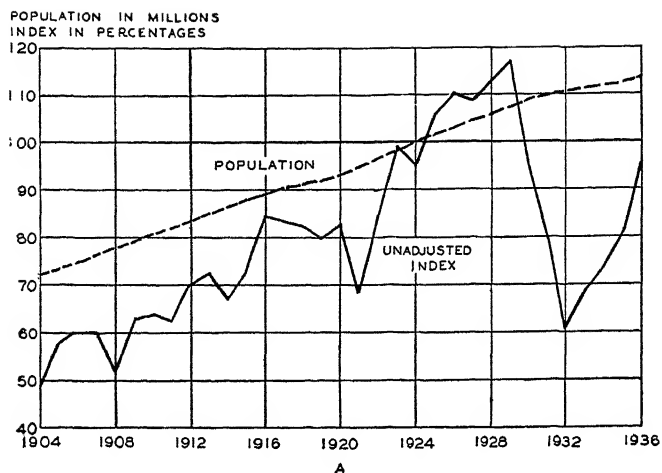
Chart 157. Combination of Population Relatives and Straight Line Values to Obtain Trend for Barron's Index of Production and Trade. (Final trend line of Section C is product of the trend lines of Sections A and B; data are from Table 91.)

cycle will be approximately equal, a number of methods are available. The moving average method has already been described.   Closely related

TABLE 91

Use of Population Estimates and Straight Line in Computing Trend Values
to Barron's Index of Trade and Production

Year (1)	Unadjusted production index (1923-1925 = 100) (2)	Population of U S relative to 1923-1925 (per cent) (3)	Index adjusted for population growth [Col 2 − Col 3] (4)	Straight line trend fitted to data of* Col 4 (5)	Final trend values [Col 3 × Col 5] (6)	Production index adjusted for trend [Col. 2 ÷ Col. 6]† (7)
1904	48 3	72 2	66 90	69 47	50 2	96.2
1905	57.8	73 6	78 53	70.72	52 0	111 1
1906	60.8	75 0	81.07	71 97	54 0	113 1
1907	60 0	76 5	78.43	73 21	56 0	107 0
1908	52 1	77 9	66 88	74 46	58 0	89 7
1909	63 2	79 2	79 80	75 70	60 0	105 5
1910	64 2	80 9	79 36	76.95	62.3	102 9
1911	62.7	82.1	76 37	78 20	64 2	97.5
1912	70 4	83 4	84 41	79.44	66 2	106 0
1913	72 4	85 0	85.18	80 69	68.6	105.4
1914	67 0	86 6	77 37	81 93	71 0	94 3
1915	72 9	87 8	83 03	83 18	73 1	99.6
1916	84 6	89 1	94 95	84 43	75 2	112.3
1917	83.7	90 3	92 69	85.67	77.4	107.9
1918	82.6	91.3	90 47	86 92	79 3	103.8
1919	80.2	91.9	87 27	88 16	81 2	99.8
1920	82 7	93.1	88 83	89 14	83.0	99.7
1921	68.3	94.8	72 05	90 66	86 0	78.8
1922	84.9	96 5	87 98	91 90	88 7	95.8
1923	99.0	98 2	100 81	93 15	91.5	107.9
1924	95.0	100 1	94.91	94 39	94.6	100.4
1925	106 1	101 6	104.43	95 64	97.1	109.2
1926	110 1	103.1	106 79	96 89	99.8	110.5
1927	108 7	104 4	104 12	98 13	102.5	106.5
1928	113 2	105.6	107 20	99 38	105.0	108.7
1929	116 8	107 2	108.96	100 65	107.9	109.4
1930	95.6	108 8	87.87	101.85	110 9	87.6
1931	79 5	109 5	72 60	103 15	113 0	71.6
1932	60 6	110.3	54.94	104 35	114 9	53.4
1933	68.3	111.1	61.48	105 65	117 3	58.2
1934	73.4	111.8	65 65	106 85	119.4	61.5
1935	81 5	112 6	72.38	108 15	121.7	66.9
1936	96.8	113 4	85.36	109 35	124.0	78.2

* Fitted to 1899-1931 data
† This column can be obtained more easily by dividing column 4 by column 5.
Source. Data furnished by *Barron's, The National Financial Weekly.*

is the following method which consists basically in obtaining one or more typical points for each cycle and connecting such points by a straight line. This method is highly subjective, and depends for its validity upon the

ability of the statistician to locate the high point and the low point of each cycle. Although annual data are used in this illustration, monthly data can be used at least equally well. In case monthly data are used, it is well, however, to smooth them by means of a 12-month centered moving average, in order to iron out seasonal variations and accidental peaks and troughs, either of which might be confused with cyclical turning points.

TABLE 92

TREND LINE BY HIGH-LOW MID-POINT METHOD FITTED TO PASSENGER AUTOMOBILE PRODUCTION, 1917–1938

Year (1)	Average monthly production (thousands of cars) (2)	Highs with interpolations (3)	Lows with interpolations (4)	High-low mid points [Average of Col 3 and Col. 4] (5)
1917	145 5 (H)	145 5 (H)	.	.
1918	78 6 (L)	149 9	78 6 (L)	114 2
1919	138 1	154 4	92 5	123.4
1920	158 8 (H)	158 8 (H)	106 3	132 6
1921	120 2 (L)	206 6	120 2 (L)	163 4
1922	189 5	254 3	168 6	211 4
1923	302 1 (H)	302 1 (H)	217 1	259 6
1924	265.5 (L)	306 5	265 5 (L)	286 0
1925	311 3	310 9	258 6	284 8
1926	315 3 (H)	315.3 (H)	251 6	283 4
1927	244 7 (L)	337 6	244.7 (L)	291 2
1928	318 0	360 0	214.7	287 4
1929	382 3 (H)	382 3 (H)	184.7	283.5
1930	232 1	375 3	154 6	265.0
1931	164.4	368.3	124 6	246 4
1932	94 6 (L)	361.3	94.6 (L)	228 0
1933	131 1	354 3	106 6	230 4
1934	181 5	347.3	118.6	233.0
1935	271 0	340 3	130 6	235.4
1936	305.8	333.3	142 7	238.0
1937	326 3 (H)	326 3 (H)	154.7	240 5
1938	166.7 (L)*		166 7 (L)	..

(H) = Cyclical High   (L) = Cyclical Low
* 1938 is taken tentatively as a cyclical low
Source Production data from United States Department of Commerce, *Survey of Current Business,* 1936 Supplement, p 147; 1938 Supplement, p. 160, February 1939, p 95

One method, which is used by the Cleveland Trust Company, is illustrated in Table 92 and Chart 158, and may be called the *high-low mid-point method.* The procedure is as follows:

(1) Determine the high point of each cycle.
(2) Connect the high points by straight lines. In Chart 158 these are light dashed lines.

(3)   Determine, by arithmetic interpolation, the values on this line for each year (see column 3 of Table 92).

(4)   Determine the low point of each cycle; connect the low points by straight lines and interpolate (see columns 2 and 4, Table 92).

(5)   Average the high and low values for each year, thus obtaining the mid-points which are shown by the heavy dashed line on Chart 158 and by the values in column 5 of Table 92.

THOUSANDS
OF CARS
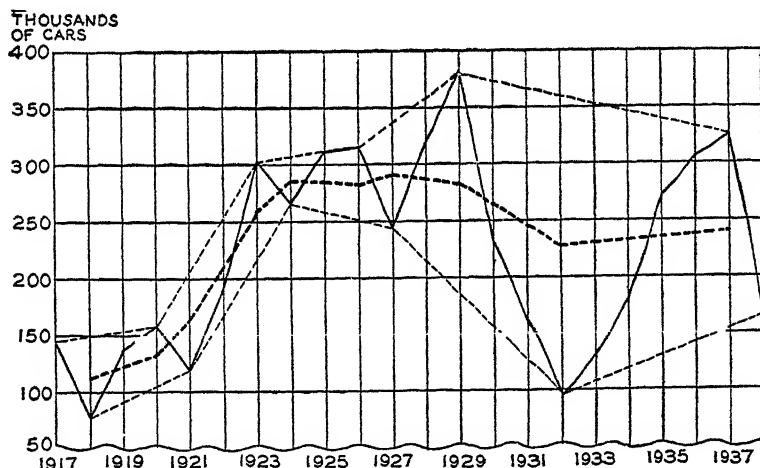
Chart 158.   Trend by High-Low Mid-Point Method Fitted to Passenger Automobile Production, 1917-1938.   (Data of Table 92.)

Variations of this method are possible, but will not be illustrated here. Thus, instead of connecting the high points of each cycle and the low points, we might connect the average value of each cycle, such points being centered at the middle of each cycle (see Babson's "X-Y line" of Chart 255 and explanation on page 814); or the points might be the averages of half cycles, running from high to low and from low to high.

All of these cyclical average methods are open to the possible objection that, first of all, the statistician decides what fluctuations he wishes to identify as cycles, and then chooses his high and low points so that the trend will go through these cycles.   These methods, therefore, while simple, require even more judgment than do moving averages, and, as is the case with the moving average method, a trend can never be up to date. Another possible objection is that the resulting curve is not very smooth. The curve may, however, be looked upon as a first approximation to be smoothed either freehand or by a mathematical curve.   Finally, it may be objected that cyclical average curves do not look like trends at all.

In reality they are not primary trends, but resemble more closely the combined primary-secondary trends, which were mentioned briefly in Chapter XIV.

Selecting the Type of Trend

(1)　As a first step the data should always be plotted and a curve fitted tentatively by inspection. The plotting should be on semi-logarithmic paper as well as arithmetic. Should the trend appear to be of a simple
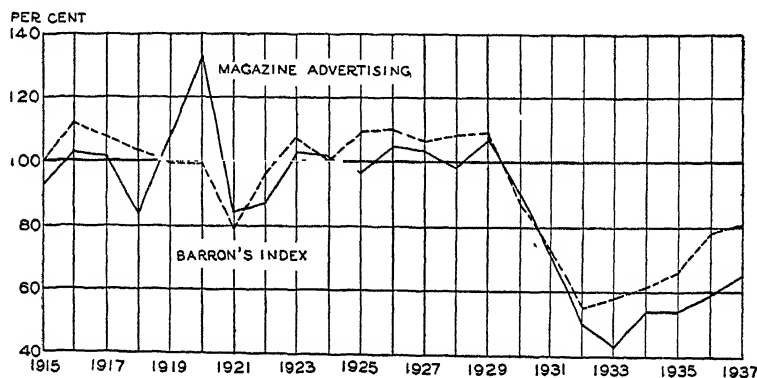
Chart 159. United States Magazine Advertising Adjusted for Trend, and Barron's Index of Production and Trade, by Years, 1915-1936. (Data of Tables 91 and 93.)

type on semi-logarithmic paper, then one of the methods to be described in the following chapter may be appropriate.

(2)　If the object of the analysis is solely to measure cyclical deviations, a reasonably smooth curve should be selected which approximately passes through the center of the different cycles. This may be accomplished by one or more straight lines, or a moving average or line of cyclical averages may be useful. A primary trend, likewise, should pass through the center of the secondary waves.

(3)　If the object is analysis of the trend itself, or prediction, a mathematical equation should be used. Such an equation is also a concise way of defining the trend.

(4)　A curve should be logical in the sense that it behaves in a manner which seems reasonable when we consider the forces affecting the series. The use of a moving average for a primary trend would seem to confess lack of complete understanding of, or hypothesis concerning, the social forces at work. A straight line also may not be logically satisfying. On the other hand, population growth is a logical explanation of part of the long term growth of many series. It is quite likely, however, that more

complicated curves, such as are described in Chapter XVI, will be required to satisfy this criterion than are described in this chapter.

(5) In case there is difficulty in deciding which among several equation types is most logical, objective tests (which will be described in Chapter XVI) may be applied which purport to show the mathematical law that the series approximates.

TABLE 93

ADJUSTMENT OF UNITED STATES MAGAZINE ADVERTISING DATA
FOR TREND, 1915–1937

(Original data and trend values in thousands of lines. Trend line fitted to data for 1915–1930 and extended from 1930 to date)

Year	Original data Y	Trend values Y_c	Per cent of trend $100(Y \div Y_c)$
1915	1,407	1,508	93
1916	1,669	1,626	103
1917	1,772	1,745	102
1918	1,547	1,864	83
1919	2,142	1,983	108
1920	2,803	2,102	133
1921	1,856	2,221	84
1922	2,030	2,340	87
1923	2,520	2,458	103
1924	2,620	2,577	102
1925	2,623	2,696	97
1926	2,958	2,815	105
1927	3,038	2,934	104
1928	3,032	3,053	99
1929	3,384	3,172	107
1930	2,984	3,290	91
1931	2,409	3,409	71
1932	1,763	3,528	50
1933	1,555	3,647	43
1934	2,027	3,766	54
1935	2,115	3,885	54
1936	2,378	4,004	59
1937	2,671	4,122	65

Source Table 89

Adjustment for Trend

It was suggested that one object of measuring trend is to measure cyclical deviations from it. If we are dealing with monthly data, a first step in obtaining such cycles may be to express the data as percentages of trend by dividing the original data by the monthly trend values (and multiplying by 100). But, in order completely to isolate cycles, we must also eliminate seasonal and irregular movements. Since it is difficult to see

what direct use can be made of data adjusted for trend alone (while, on the other hand, seasonally adjusted data are in common use), it is customary to eliminate the seasonal first and then the trend.   Consequently, in this chapter, adjustment of monthly data for trend is not made.   However, the magazine advertising annual data are divided[7] by their trend values, giving the annual values of the cyclical movements.   These cyclical relatives based on annual data are rough measures and can be used in comparison with other annual data similarly adjusted.   The process of computation is shown in Table 93, and the resulting cyclical relatives are plotted in Chart 159.   This series shows much the same peaks and troughs as most economic series, but indicates a tendency to lag behind general business.   Note that cyclical troughs occurred in 1925, 1928, and 1933, rather than in 1924, 1927, and 1932.

Selected References

E. C. Bratt: *Business Cycles and Forecasting*, Chapter III; Business Publications, Inc., Chicago, 1937.   Mainly a consideration of economic factors.

R. E. Chaddock· *Principles and Methods of Statistics*, pages 306–336; Houghton Mifflin Co., Boston, 1925.   Trend fitting is considered as a special application of correlation analysis.

F. E. Croxton and D. J. Cowden: *Practical Business Statistics*, Chapter XV; Prentice-Hall, Inc , New York, 1934.

F. C. Mills: *Statistical Methods Applied to Economics and Business* (Revised Edition), pages 231–253; Henry Holt and Co., New York, 1938.   Moving averages are fitted to hypothetical data, clearly illustrating the principles involved.

E. C. Rhodes: *Elementary Statistical Methods*, pages 211–233; George Rutledge and Sons, London, 1933.   Moving averages.

C. H Richardson: *An Introduction to Statistical Analysis*, Chapter VI; Harcourt, Brace and Co., New York, 1934.   Applies the method of least squares and the method of moments to fitting linear trends.

---

[7] We could subtract the trend rather than divide by it.   This would give absolute rather than relative deviations.   For most purposes, however, it is more useful to know whether the variations are large relative to some logical base, such as the trend. Thus, a deviation of 50 is ten times as important when judged with respect to a trend value of 200 than it is when compared with a trend value of 2,000.

CHAPTER XVI

OTHER TREND TYPES

In Chapter XV only the simplest type of trend equation was discussed—the straight line. Frequently it is the case that for short periods of time the straight line gives a reasonably good fit. When viewed over a long period of time, however, many series do not appear to follow so simple a "law." Usually the slope is gradually changing; even the change in the slope may be changing. It is apparent, for instance, that the growth of rayon consumption since 1919 is not adequately described by a straight line.[1] Even a casual inspection of Chart 160 reveals that any adequate trend line for these data must have at least one bend in it.

It is the object of this chapter to describe the properties of several such equations, to give directions for their fitting, and to explain in somewhat greater detail how to select from among the numerous types at the disposal of the statistician.

Weighted Moving Averages

Before describing more equation types, however, it is well to note that smoother and more flexible results can be obtained by the introduction of weights into moving averages, than can be had by the use of ordinary unweighted moving averages. A centered 2-year moving average, it was noted, could also be thought of as a weighted 3-year moving average in which the middle year is given a weight of two. A system of weights often used is known as binomial. Thus, $(a + b)^2 = a^2 + 2ab + b^2$, giving the weights 1, 2, 1. Binomially, weights for a 5-year moving average can be obtained from $(a + b)^4 = a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4$, the weights being 1, 4, 6, 4, 1. For 7 years they are 1, 6, 15, 20, 15, 6, 1.

It would be laborious to compute directly a 7-year binomially weighted

---

[1] Although data are available as far back as 1911, it does not seem advisable in this instance to fit one curve to the whole period. Judging from plotted data, it appears that the World War so interrupted the early growth of the industry that it practically had a fresh start in 1919.

421

moving average, but a short cut is possible. Compute first a weighted
3-year moving average; then take a 3-year weighted moving average of
the result, and another weighted 3-year moving average of this result.[2]
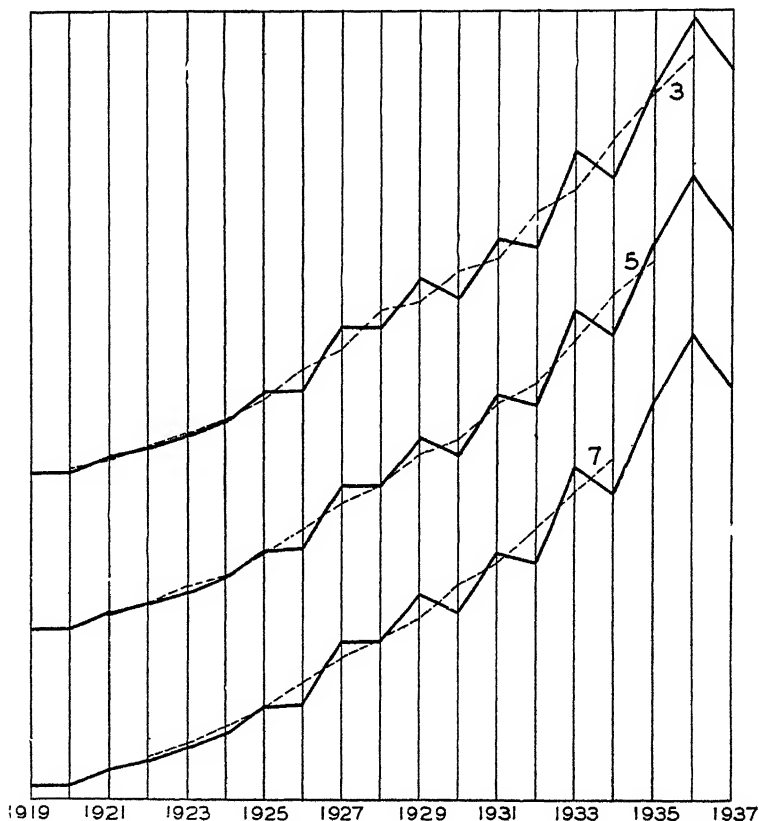
Chart 160. Binomially Weighted Moving Average Trends Fitted to United States
Rayon Consumption, 1919-1937. (For purposes of comparison the different curves
have been plotted close together on the same chart instead of on separate charts. Each
curve is plotted to the same vertical scale, but at a different level. This arrangement
is sometimes referred to as a multiple axis chart. Data of Table 94.)

This procedure is followed in Table 94. In general language, for a bino-
mial of N terms, take a binomially weighted 3-term moving average suc-
cessively $\dfrac{N-1}{2}$ times.

It is very easy to compute a weighted average at one operation on a

---

[2] The logic of this procedure is perhaps most easily understood by a careful study of
the table on page 423, the construction of which is self-explanatory.

Derivation of Binomial Weights for Moving Average

Year	Value for year	Year				Weight system, centered on year 2 [sum of columns 3–6]	Year				Weight system, centered on year 3 [sum of columns 8–11]	Year				Weight system, centered on year 4 [sum of columns 13–16]
		1	2 (double weight)		3		2	3 (double weight)		4		3	4 (double weight)		5	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
1	Y_1	1	·			$1Y_1$	1				$1Y_1$	1				$1Y_1$
2	Y_2	·	1	1		$2Y_2$	2	1	1		$4Y_2$	4	1	1		$6Y_2$
3	Y_3				1	$1Y_3$	1	2	2	1	$6Y_3$	6	4	4	1	$15Y_3$
4	Y_4							1	1	2	$4Y_4$	4	6	6	4	$20Y_4$
5	Y_5									1	$1Y_5$	1	4	4	6	$15Y_5$
6	Y_6												1	1	4	$6Y_6$
7	Y_7														1	$1Y_7$
Sum of weights	··	·	·	·	·	4	·				16					64

calculating machine of the Monroe or Marchant type. The 3-year moving average (11,620, in Table 94) is obtained as follows·

(1) Set machine for multiplication.
(2) Put in keyboard . . 9,291; depress plus bar; clear keyboard
(3) Put in keyboard 8,718; depress plus bar twice; clear keyboard.
(4) Put in keyboard . . 19,751; depress plus bar; clear keyboard
(5) Total in lower dial is. 46,478
(6) Put in keyboard . 46,478, clear both dials
(7) Multiply by 25, the reciprocal of 4, obtaining 11,620

TABLE 94

COMPUTATION OF BINOMIALLY WEIGHTED ⁻-YEAR MOVING AVERAGE OF UNITED STATES
RAYON CONSUMPTION, 1919–1937

(Thousands of pounds)

Year	Consumption	Moving average		
		3-year	5-year	7-year
1919	9,291	.		.
1920	8,718	11,620		
1921	19,751	18,242	18,389	.
1922	24,747	25,451	25,542	25,826
1923	32,558	33,026	33,833	34,274
1924	42,243	43,830	43,886	44,366
1925	58,277	54,857	55,860	56,705
1926	60,630	69,896	71,214	71,962
1927	100,048	90,207	89,558	89,226
1928	100,101	107,924	106,574	105,651
1929	131,448	120,241	119,898	119,642
1930	117,968	131,186	132,199	133,066
1931	157,360	146,182	147,970	149,253
1932	152,041	168,331	168,872	169,375
1933	211,883	192,644	191,786	192,431
1934	194,771	213,525	217,281	218,440
1935	252,676	249,429	247,412	..
1936	297,594	277,265	..	...
1937	261,195		.	...

Source: See Table 80.

It is well to plot the original data as well as the smoothed data after each successive smoothing as in Chart 160. This will accomplish the two-fold object of bringing to light any large errors in computation, and telling when the data have been sufficiently smoothed.

The greater smoothness of weighted moving averages is due to the fact that an item exerts slight influence on the final average when it is first included, but it gradually grows more powerful, and then as gradually dwindles in importance until it finally disappears. This will be clear if

we examine the weight pattern of a binomial of 15 items, which can be obtained by seven smoothings of original data by weighted 3-item moving averages. The weight pattern is: 1; 14; 91; 364; 1,001; 2,002; 3,003; 3,432; 3,003; 2,002; 1,001; 362; 91; 14; 1. Thus any member, when it appears for the first or last time, exerts only $\frac{1}{16,384}$ of the influence exerted by all the numbers (16,384 being the sum of all the frequencies in the weight pattern), while in a simple moving average the initial or terminal influence is $\frac{1}{15}$). On the other hand, greater flexibility is obtained by the fact that the maximum influence obtained by any one number in a binomial 15 is $\frac{3,432}{16,384}$, as compared with $\frac{1}{15}$ for the unweighted average. Chart '61 is a diagram of the above weight pattern. A binomial weighting sys-
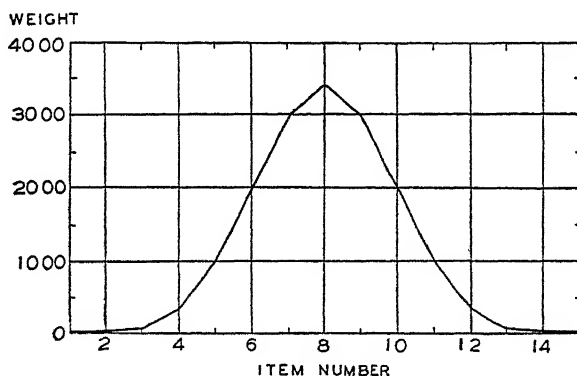
Chart 161.   Weight Diagram of a 15-Item Binomially Weighted Moving Average.   The weights total 16,384.

tem is but one type of weight pattern, some types even introducing negative weights in certain portions. In general it may be said that the smoother the weight pattern, the smoother the resulting trend. Since a binomial weight pattern is very smooth, so also is the trend resulting from the use of it.

Because of the irregularity of cyclical movements it is usually not readily apparent what length of moving average to use. It is therefore convenient merely to smooth the data by successive moving averages until the desired smoothness is obtained. Seven years seems sufficient for the rayon data. One objection to a weighted moving average similar to those discussed is that it requires so many terms to smooth out undesirable minor wave-like movements, thus removing so many end values from the trend. It seems

not unreasonable to contend, therefore, that for most purposes sufficiently good results will be obtained by using a smaller number of years in a simple moving average, and then smoothing the results by a small term binomial moving average, or on a chart freehand. Several other objections to moving average trends, regardless of the weighting system, were discussed in Chapter XV. It is worth repeating that no expression for the trend is available when we use moving averages.

Simple Polynomials

This family of curves has as its most elementary type the straight line, which, it will be remembered, has two constants. As will be explained, additional constants introduce one or more bends into the curve. Below are given the equation types of the five simplest varieties:

First degree (straight line) $Y_c = a + bX$
Second degree (parabola)   .   . $Y_c = a + bX + cX^2.$
Third degree (cubic)   . $Y_c = a + bX + cX^2 + dX^3.$
Fourth degree (quartic)   .. $Y_c = a + bX + cX^2 + dX^3 + eX^4$
Fifth degree (quintic)   .   . . $Y_c = a + bX + cX^2 + dX^3 + eX^4 + fX^5.$

Fourth or fifth degree curves, which may change in slope from positive to negative direction, or from negative to positive direction, respectively three and four times, hardly coincide with the concept of primary trend as set forth previously, and only the second degree curve will be explained in detail here.

Second degree curve. This curve is but one degree more complicated than a straight line. It differs in that its slope is continually changing in such a way that the curve has one bend. If a sufficient number of X values are included, it is inclined positively in one part and negatively in another. Eight of these curves are shown in Chart 162.

The mechanics of determining the Y_C values are not difficult, and will be illustrated for the curve shown in section 1 of the chart referred to. The values of a, b, and c are taken to be 5, 2, and $-.3$ respectively; and the equation is therefore

$$Y_C = 5 + 2X - .3X^2.$$

When X is 0,

$$Y_c = 5 + 2(0) - .3(0)^2 = 5.$$

When X is -4,

$$Y_c = 5 + 2(-4) - .3(-4)^2 = 5 - 8 - 4.8 = -7.8.$$

In like manner other values may be substituted, with the results that are tabulated at the right of this curve in Chart 162.

The meaning of the constants a, b, and c as applied to second degree curves may now be summarized: a indicates the Y_C value when $X = 0$ b indicates the amount and direction of the slope at the point where
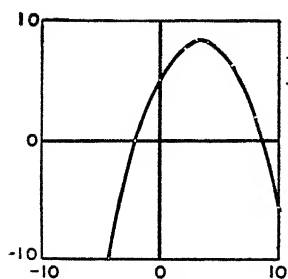
X	Y_c
-4	-7 8
-2	.4
0	5 0
2	7 8
3	8 3
4	8 2
6	6 2
8	1.8
10	-5 8

(1) $Y_c = 5+2X-.3X^2$

X	Y_c
-3	-9 7
-1	-3 3
0	-1 0
2	1 8
3	2 3
4	2 2
5	1.5
7	1.7
9	-7 3

(2) $Y_c = -1+2X-.3X^2$

X	Y_c
-10	3.0
-8	1 4
-6	.6
-5	5
-4	6
-2	1 4
0	3 0
2	5 4
4	8 6

(3) $Y_c = 3 + X + 1X^2$

X	Y_c
-4	8 6
-2	5 4
0	3 0
2	1 4
4	6
5	5
6	.6
8	1.4
10	3.0

(4) $Y_c = 3 - X + 1X^2$

X	Y_c
-10	9 00
-7	4 95
-4	1.80
-2	.20
0	-1.00
3	-2.05
5	-2 25
7	-2.05
10	-1 00

(5) $Y_c = -1 - 5X + .05X^2$

X	Y_c
-10	-1.00
-8	20
-6	.20
-4	20
-2	-.20
0	-1 00
3	-2 95
6	-5 80
9	-9 55

(6) $Y_c = -1 - .5X - .05X^2$

X	Y_c
-10	0
-8	-.80
-6	-1 20
-4	-1.20
-2	-80
0	0
3	1 95
6	4 80
10	10 00

(7) $Y_c = .5X + .05X^2$

X	Y_c
-10	-5.00
-7	-2 45
-4	-80
-2	20
0	0
2	-20
4	-80
7	-2.45
10	-5 00
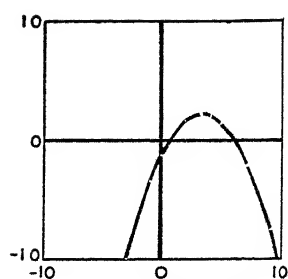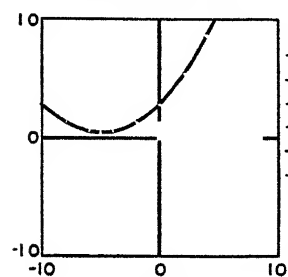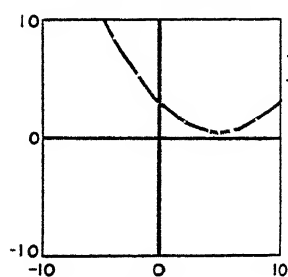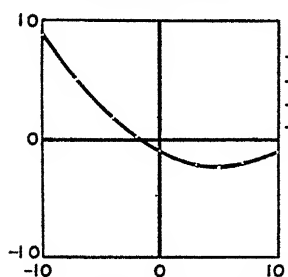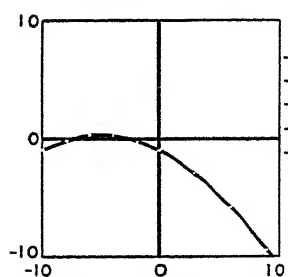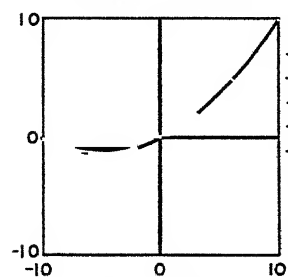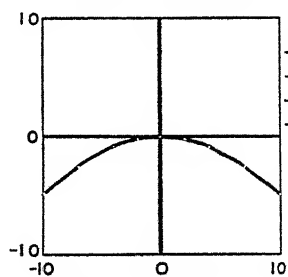
(8) $Y_c = -.05X^2$

Chart 162.   Second Degree Equations and Curves.

427

$X = 0$; $2c$ indicates the amount of change in the slope per unit of X, and whether this curvature is such as to make it slope upward or downward when the value of X is taken as large and positive.

TABLE 95

SECOND DEGREE CURVE FITTED TO UNITED STATES RAYON CONSUMPTION

(Thousands of pounds)

Year	X	Con-sumption Y	XY	X^2Y	Computation of trend values			
					X^2	$a + bX$	cX^2	Trend value Y_C
1919	−17	9,291	−157,947	2,685,099	289	−44,795 5	54,397 1	9,602
1920	−15	8,718	−130,770	1,961,550	225	−28,020 0	42,350.7	13,331
1921	−13	19,751	−256,763	3,337,919	169	−13,244.6	31,810.1	18,566
1922	−11	24,747	−272,217	2,994,387	121	2,530 9	22,775.2	25,306
1923	− 9	32,558	−293,022	2,637,198	81	18,306 3	15,246 2	33,552
1924	− 7	42,243	−295,701	2,069,907	49	34,081 8	9,223 0	43,305
1925	− 5	58,277	−291,385	1,456,925	25	49,857 2	4,705.6	54,563
1926	− 3	60,630	−181,890	545,670	9	65,632 6	1,694.0	67,327
1927	− 1	100,048	−100,048	100,048	1	81,408 1	188 2	81,596
1928	1	100,101	100,101	100,101	1	97,183.5	188 2	97,372
1929	3	131,448	394,344	1,183,032	9	112,959 0	1,694 0	114,653
1930	5	117,968	589,840	2,949,200	25	128,734 4	4,705 6	133,440
1931	7	157,360	1,101,520	7,710,640	49	144,509 9	9,223 0	153,733
1932	9	152,041	1,368,369	12,315,321	81	160,285.3	15,246 2	175,532
1933	11	211,883	2,330,713	25,637,843	121	175,060 8	22,775 2	198,836
1934	13	194,771	2,532,023	32,916,299	169	191,836 2	31,810 1	223,646
1935	15	252,676	3,790,140	56,582,100	225	207,611.6	42,350 7	249,962
1936	17	297,594	5,059,098	86,004,666	289	223,387 1	54,397 1	277,784
Total		1,972,105	15,286,405	243,457,905			.	

Source See Table 80

$$\text{II} \quad 15,286,405 = 1,938b$$
$$b = 7,887.722.$$
$$\text{I.} \quad 1,972,105 = 18a + 1,938c.$$
$$\text{III} \quad 243,457,905 = 1,938a + 374,034c.$$
$$(\text{I} \times 107.66667) \quad 212,329,978 = 1,938a + 208,658.01c$$
$$\text{III.} \quad \underline{243,457,905 = 1,938a + 374,034.00c}$$
$$31,127,927 = \qquad\qquad 165,375.99c$$
$$c = 188.22519.$$
$$\text{I.} \quad 1,972,105 = 18a + 1,938(188.22519)$$
$$18a = 1,607,324\ 58$$
$$a = 89,295\ 810$$

Check (III):
$$243,457,905 = 1,938(89,295\ 810) + 374,034(188.22519)$$
$$= 243,457,900.5.$$

Trend equation:
$$Y_C = 89,295.810 + 7,887.722X + 188.22519X^2.$$

Origin, 1927–1928. X units ½ year.

Since there are three unknowns or constants, three normal equations (each with three constants) are required for a second degree curve:

$$\text{I.} \qquad \Sigma Y = Na + b\Sigma X + c\Sigma X^2.$$
$$\text{II.} \qquad \Sigma XY = a\Sigma X + b\Sigma X^2 + c\Sigma X^3.$$
$$\text{III.} \qquad \Sigma X^2Y = a\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4.$$

When we are dealing with time series, however, and the origin is taken at the middle of the period, the odd powers of X, of course, total zero, and the equations become:

$$\text{I.} \qquad \Sigma Y = Na + c\Sigma X^2.$$
$$\text{II.} \qquad \Sigma XY = b\Sigma X^2.$$
$$\text{III.} \qquad \Sigma X^2Y = a\Sigma X^2 + c\Sigma X^4.$$

The only values which must be computed are therefore ΣY, ΣXY, ΣX^2Y, since ΣX^2 and ΣX^4 can be obtained from Appendix M or N.

The computation of these values and the solution of the equations are shown in Table 95. It is to be noted that only equations I and III need to be solved simultaneously, since equation II has only one unknown (b) and equations I and III have two each (a and c). A word of explanation concerning the simultaneous solution of equations I and III is advisable. Dividing the coefficient of a in equation I into that of a in equation III gives 107.66667 Therefore, in order to cancel out a, equation I is multiplied by 107.66667. Then equation I is subtracted from equation III, the result being $31{,}127{,}927 = 165{,}375.99c$, from which c is calculated to be 188.22519. This value of c is now substituted in equation I, and a value for a obtained, (89,295.810).

As a check on the accuracy of the solution, the values of a and c are substituted in equation III, which gives an agreement to eight digits. This does not prove that the values originally substituted in the normal equations were correct, but, granting their accuracy and barring counterbalancing errors, there has been no important mistake in the solution of equations I and III. It is desirable that this check be made, even though no check for b is available in this setup. The equation may now be stated: $Y_c = 89{,}295.81 + 7{,}887.722X + 188.22519X^2$, with origin between 1927 and 1928, and X units one-half year. The computation of the Y_c values is shown in the last four columns of Table 95, and the values are plotted in Chart 163. It is apparent that the fit is excellent. The curve cuts under all the prosperity peaks except 1927, which was an unusually low peak, perhaps on account of the mild general depression in the United States that year. Note that 1936 is a little unusual also. According to precedent that year should have been a depression year, and in fact the *percentage rate* of increase from 1935 to 1936 was less than from 1934 to 1935. Nevertheless, this year was not below the trend line. Again it

may be that the general business pickup in 1936 served to make the rayon depression less severe than usual.

Third degree curve. By adding one more constant to the equation, we are enabled to put one more bend into our trend curve. In Chart 164 are shown two such equations and curves. It should be noted that a straight line has only one slope, a second degree curve slopes in a positive
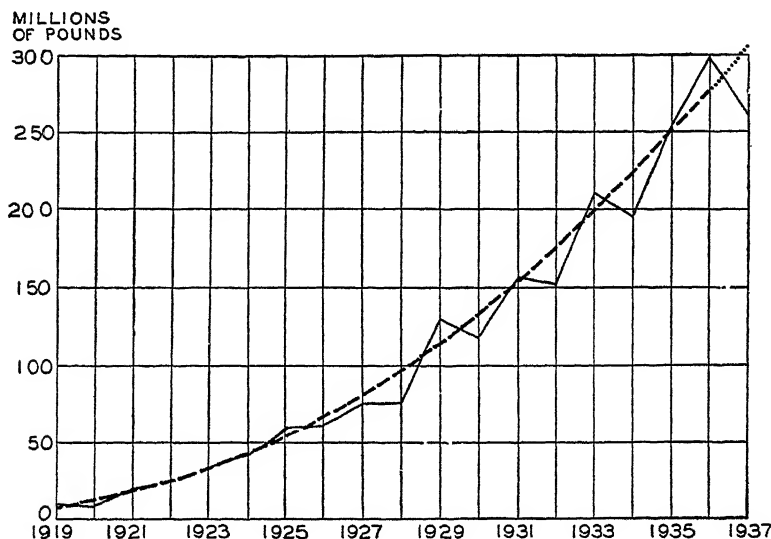
Chart 163. Second Degree Curve Fitted to United States Consumption of Rayon 1919-1936, and Trend Extension Through 1937. (Data of Table 95.)

direction at one stage and in a negative direction at another, while a third degree curve includes three directions of slope.

Four normal equations are required for a third degree curve:

$$\text{I.} \quad \Sigma Y = Na + b\Sigma X + c\Sigma X^2 + d\Sigma X^3.$$
$$\text{II.} \quad \Sigma XY = a\Sigma X + b\Sigma X^2 + c\Sigma X^3 + d\Sigma X^4.$$
$$\text{III.} \quad \Sigma X^2Y = a\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4 + d\Sigma X^5.$$
$$\text{IV.} \quad \Sigma X^3Y = a\Sigma X^3 + b\Sigma X^4 + c\Sigma X^5 + d\Sigma X^6.$$

Again, if the X origin is taken at the middle of the period, the odd powers of X will cancel, leaving these equations:

$$\text{I.} \quad \Sigma Y = Na + c\Sigma X^2.$$
$$\text{II.} \quad \Sigma XY = b\Sigma X^2 + d\Sigma X^4.$$
$$\text{III.} \quad \Sigma X^2Y = a\Sigma X^2 + c\Sigma X^4.$$
$$\text{IV.} \quad \Sigma X^3Y = b\Sigma X^4 + d\Sigma X^6.$$
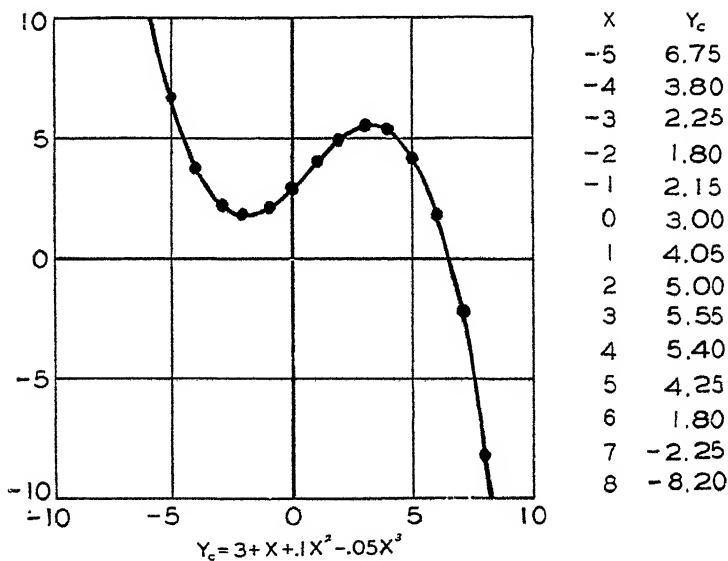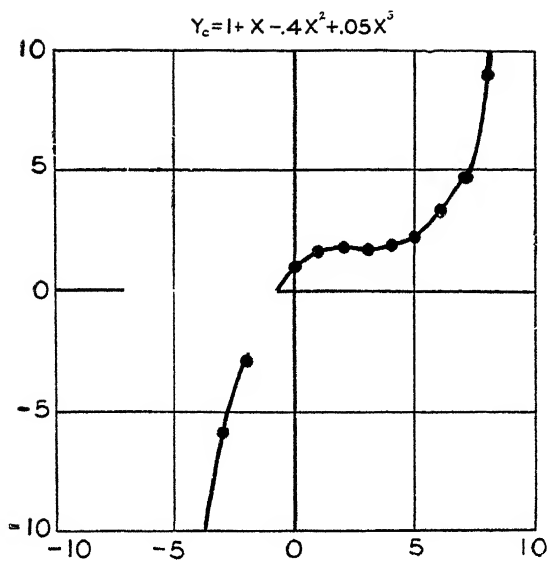
Chart 164. Third Degree Equations and Curves.

431

To obtain the trend equation, equations I and III, containing a and c, must now be solved simultaneously, and likewise equations II and IV, which contain b and d. Only one new column of figures in addition to those of Table 95 need be computed; it is the column necessary to obtain $\Sigma X^3 Y$. Furthermore, equations I and III are exactly the same as for the second degree curve; consequently the values of a and c will also be the same.

A fourth degree curve, if origin is taken at the middle of the period, necessitates these normal equations:

$$\text{I.} \quad \Sigma Y = Na + c\Sigma X^2 + e\Sigma X^4.$$
$$\text{II.} \quad \Sigma XY = b\Sigma X^2 + d\Sigma X^4.$$
$$\text{III.} \quad \Sigma X^2 Y = a\Sigma X^2 + c\Sigma X^4 + e\Sigma X^6.$$
$$\text{IV.} \quad \Sigma X^3 Y = b\Sigma X^4 + d\Sigma X^6.$$
$$\text{V.} \quad \Sigma X^4 Y = a\Sigma X^4 + c\Sigma X^6 + e\Sigma X^8.$$

Such a setup permits equations II and IV to be solved simultaneously, and equations I, III, and V. Some persons prefer, when solving three or more equations simultaneously, to employ a systematic, self-checking procedure, such as the Doolittle method explained in Chapter XXIII.

Empirical test of data. An additional property of this family of curves is:

(1) The first differences of a first degree curve are constant.
(2) The second differences of a second degree curve are constant.
(3) The third differences of a third degree curve are constant.
(4) The nth differences of an nth degree curve are constant.

The first and second differences of the trend values of the second degree fit are shown in Table 96. As can easily be observed, a first difference is merely the difference between any number and the number preceding it; while a second difference is the same thing with respect to first differences. The slight discrepancies in the fourth digit are, of course, due to the rounding of the trend values.

This knowledge concerning the properties of the potential series can be used to test any time series data for trend type. If the data are reasonably regular, it is possible to take successive differences of the data until the differences most nearly approaching constancy are obtained. It is not often enlightening, however, to so test the original data, for the successive differences of the cyclical changes would be so pronounced that little could be discovered about the underlying trend. When this is true, it is probably better first to approximate a trend by one of the flexible methods which do not involve the obtaining of an equation, and then to take the differences of the trend. Such an approximation might be based on a freehand smoothing of the high-low mid-point or some other cyclical average method, or the moving average method. Although this is apparently

an objective method, in practice it is usually very difficult to determine which column of differences is most nearly constant.

TABLE 96

FIRST AND SECOND DIFFERENCES OF SECOND DEGREE TREND
VALUES FOR UNITED STATES RAYON CONSUMPTION DATA.
1919–1936

(Thousands of pounds)

Year	Trend values	First differences	Second differences
1919	9,772		.
1920	13,481	3,709	. .
1921	18,696	5,215	1,506
1922	25,416	6,720	1,505
1923	33,642	8,226	1,506
1924	43,375	9,733	1,507
1925	54,613	11,238	1,505
1926	67,357	12,744	1,506
1927	81,606	14,249	1,505
1928	97,362	15,756	1,507
1929	114,623	17,261	1,505
1930	133,390	18,767	1,506
1931	153,663	20,273	1,506
1932	175,442	21,779	1,506
1933	198,726	23,284	1,505
1934	223,516	24,790	1,506
1935	249,812	26,296	1,506
1936	277,614	27,802	1,506

Source Table 95

Orthogonal polynomials. A disadvantage of polynomial equations of the type described is that each additional constant added to the equation requires that some of the constants previously obtained be abandoned and new constants computed to take their place. Thus, a second degree curve uses the same value for b as a straight line, but requires a different value for a; a third degree curve uses the same values for a and c as a second degree curve, but requires a new value for b; a fourth degree curve uses the same values for b and d as a third degree curve, but new values must be calculated for a and c; and so on. *Orthogonal polynomial* equations involve a transformation of such a nature that, as new constants are added, the old constants remain the same. Such equations are very convenient to use, since we merely build up our equation by adding new constants until a satisfactory fit is obtained and simultaneous solution of equations is avoided. There is thus no lost motion, and the labor involved becomes progressively less than that required to fit a curve by the ordinary

method for equations of third degree and higher. The trend values obtained by the two methods are exactly the same.

Although the labor required for fitting is modest, the theory of orthogonal polynomials is beyond the scope of this text, and will not be explained here. Whereas the ordinary third degree polynomial is of the type

$$Y_C = a + bX + cX^2 + dX^3,$$

the orthogonal polynomial is

$$Y_C = A + BX_1 + CX_2 + DX_3.$$

In working with orthogonal polynomials, the X origin is conveniently taken at the middle so that $\Sigma X = 0$. If N is odd, the X values are taken as $\cdots -3, -2, -1, 0, +1, +2, +3 \cdots$ in the usual fashion; if N is even, they are taken as $\cdots -2.5, -1.5, -.5, +.5, +1.5, +2.5 \cdots$. The variables $X_1, X_2, X_3 \cdots$ are derived from the moments of the X series. In form easy to use, these are:

$$X_1 = X.$$

$$X_2 = X_1^2 - \frac{N^2 - 1}{12}.$$

$$X_3 = X_1^3 - \frac{3N^2 - 7}{20} X_1.$$

$$X_{(r+1)} = X_1 X_r - \frac{r^2(N^2 - r^2)}{4(4r^2 - 1)} X_{(r-1)}.$$

N is, as usual, the number of items in the series—the number of years or months—and r is the degree of the polynomial. Each of these equations is worked out, and in the computation table there will be column headings for X_1, X_2, and X_3. The constants A, B, C, and D will be obtained as follows:

$$A = \frac{\Sigma Y}{N}.$$

$$B = \frac{12}{N(N^2 - 1)} \Sigma X_1 Y.$$

$$C = \frac{180}{N(N^2 - 1)(N^2 - 4)} \Sigma X_2 Y.$$

$$D = \frac{2800}{N(N^2 - 1)(N^2 - 4)(N^2 - 9)} \Sigma X_3 Y.$$

$$\text{Coefficient of } X_r = \frac{(2r)! \, (2r+1)!}{(r!)^4 N(N^2 - 1)(N^2 - 4) \cdots (N^2 - r^2)} \Sigma X_r Y.$$

In obtaining the trend values, the constants are multiplied by X_1, X_2, and X_3 instead of X, X^2, and X^3. For the theory of orthogonal polyno-

mials the reader is referred to R. A. Fisher, *Statistical Methods for Research Workers* (7th Edition), pp. 148–155. Fisher also explains a short cut method of fitting, which consists almost entirely of successive additions.

Use of Logarithms

Straight line equation. It is quite apparent from inspection of Chart 165 that the curved trend line fits the Japanese industrial production data much better than would a straight line. It is not always necessary, however, to fit an equation with three constants in order to obtain a curve with one upward bend in it. The exponential curve, $Y_C = ab^X$, which represents a trend with constant rate of increase, is such a curve. In logarithmic form it becomes $\log Y_C = \log a + X \log b$. Therefore, it is usually advisable first to plot the data on semi-logarithmic paper (that is, with logarithmic vertical scale) to see if the trend seems to straighten out.[3] This is done in section B of this chart, and the straight line appears to be a reasonable fit, indicating a constant percentage rate of growth, as contrasted with a curve of the type $Y_C = a + bX + cX^2$, which indicates a type of growth that is increasing absolutely but declining relatively. In order to fit a line that is straight on semi-logarithmic paper, it is necessary only to fit a straight line to the logarithms of the Y values. Thus in Table 97, column 3, are recorded the logarithms of the production index. From this point on, the procedure for fitting the curve is the same as for any straight line. The normal equations for $\log Y_c = \log a + X \log b$, with origin at the mean of the X values (middle year), are:

$$\text{I.} \quad \Sigma \log Y = N \log a.$$
$$\text{II.} \quad \Sigma X \log Y = \log b \ \Sigma X^2.$$

The solution of these equations gives the following values:

$$\log a = 2.0158528,$$
$$\log b = .0264071,$$

and the trend equation is

$$\log Y_C = 2.015853 + .026407X,$$

---

[3] If a series does not become a straight line when plotted on semi-logarithmic paper, it is sometimes possible to accomplish this result by subtracting, algebraically, a correction factor from each observation. After the trend is fitted, the correction factor is added, algebraically. In order to obtain the correction factor, we first divide the data into three equal groups of years and compute partial totals $(\Sigma_1 Y; \Sigma_2 Y; \Sigma_3 Y)$ for each section. The correction factor is

$$k = \frac{1}{n} \left[\frac{(\Sigma_1 Y)(\Sigma_3 Y) - (\Sigma_2 Y)^2}{\Sigma_1 Y + \Sigma_3 Y - 2\Sigma_2 Y} \right].$$

In this expression n is the number of observations in a group. Compare with Frederick C. Mills, *Statistical Methods*, pp. 667–671, Henry Holt and Company, New York, 1938 (Revised Edition).
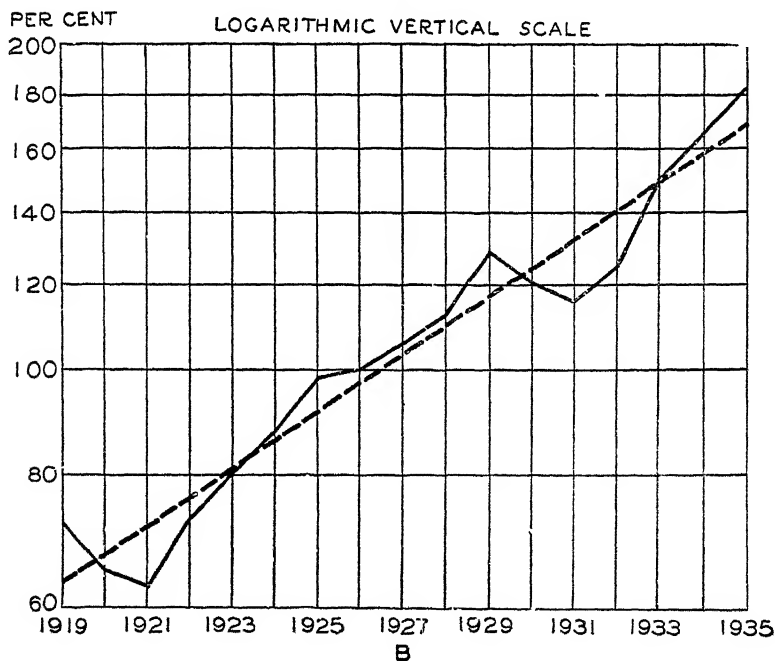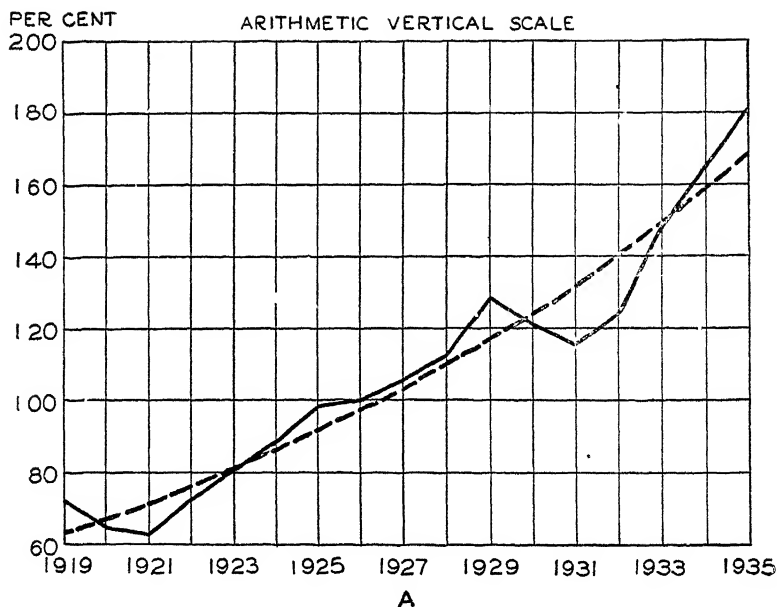
Chart 165. Japanese Industrial Production and Straight Line Trend Fitted to Logarithms, 1919-1935: A. Arithmetic Vertical Scale; B. Logarithmic Vertical Scale. (Data of Table 97.)

with origin at 1927, and X units of one year. In column 6 are recorded the log Y_C values. In order to obtain the arithmetic trend values corresponding to the original data, it is necessary only to look up the antilogarithms of the log Y_C values. The trend line is plotted in both sections of Chart 165.

Since we multiply numbers by adding their logarithms and raise them to

TABLE 97

COMPUTATION OF STRAIGHT LINE TREND TO LOGARITHMS OF JAPANESE INDUSTRIAL PRODUCTION, 1919–1935

Year (1)	Production (1926 = 100) Y (2)	log Y (3)	X (4)	X log Y (5)	Computation of trend values	
					log Y_C (6)	Y_C (7)
1919	72 51	1 860398	−8	−14 883184	1 804597	63.77
1920	65 07	1 813381	−7	−12 693667	1 831004	67.76
1921	63 03	1 799547	−6	−10 797282	1 857411	72 01
1922	73 07	1 863739	−5	− 9 318695	1 883818	76.53
1923	80 61	1 906389	−4	− 7 625556	1.910225	81.33
1924	88.54	1.947140	−3	− 5 841420	1 936632	86.42
1925	98 33	1 992686	−2	− 3 985372	1.963039	91.84
1926	100.00	2 000000	−1	− 2.000000	1 989446	97.60
1927	106 40	2.026942	0	0	2 015853	103.72
1928	113.20	2 053846	1	2 053846	2 042260	110.22
1929	128 70	2 109579	2	4.219158	2 068667	117 13
1930	120.70	2 081707	3	6 245121	2 095074	124.47
1931	116 00	2 064458	4	8 257832	2.121481	132.28
1932	124 5	2 095169	5	10.475845	2.147888	140.57
1933	149 6	2 174932	6	13.049592	2.174295	149 38
1934	165 5	2.218798	7	15.531586	2.200702	158.75
1935	182 3	2 260787	8	18.086296	2.227109	168.62
Total	..	34 269498	..	10.774100	...	.

Source Standard Statistics Company, Inc , *Standard Trade and Securities, Basic Statistics*, Vol 80. June 5, 1936, Other countries, p I–19.

LOGARITHMIC FORM	NATURAL FORM
I. 34 269498 = 17 log a	
log a = 2 0158528.	a = 103.718.
II. 10 774100 = 408 log b	
log b = .0264071.	b = 1.06269.

Trend equation:

log Y_C = 2 015853 + .026407X.

Origin, 1927; X units, one year.

Trend equation:

Y_C = 103 718(1.06269)^X

Origin, 1927; X units, one year.

a power by multiplying the logarithm of the number by that power, we may change

$$\log Y_C = \log a + X \log b$$

to

$$Y_C = ab^X.$$

Below Table 97 it is shown that $a = 103.718$ and $b = 1.06269$. Therefore, the trend equation

$$\log Y_C = 2.015853 + .026407X$$

may be written

$$Y_C = 103.718 \ (1.06269)^X.$$

The advantage of the equation in this form is that it shows 103.718 to be the trend for 1927, and that Japanese industrial production has a normal annual growth of 6.269 per cent. Incidentally it might be noted that 103.718 is the geometric mean of the series.
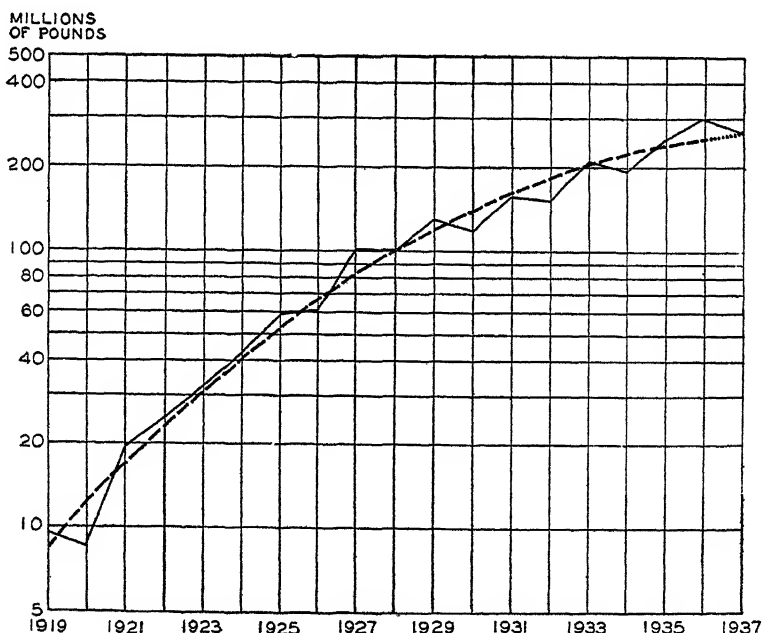
Chart 166. United States Rayon Consumption and Second-Degree Curve Fitted to Logarithms, 1919-1936, and Trend Extension Through 1937. (Data of Tables 80 and 98.)

Since the geometric mean is always a little smaller than the arithmetic mean, and since the sum of the squares of the deviations of the logarithms (rather than the original data) is at a minimum for this trend, the sum of the deviations above the trend line of Chart 165A is slightly greater than the sum of those below it. This is possibly a slight objection to this type of trend. On the other hand, the deviations on either side of the line in section B do cancel. Furthermore, use of logarithms equalizes the importance of the cycles in the early years, when the data are of small absolute size, with those of the later years, when the cycles are larger

TABLE 98

Computation of Second Degree Trend to Logarithms of United States Rayon Consumption, 1919–1936

(Thousands of pounds)

Year	Consumption Y	log Y	X	X log Y	X²	X² log Y	Computation of trend values			
							a + bX	cX²	log Yc	Yc
1919	9,291	3.968062	−17	−67.4.7054	289	1146.769918	4.280112	−.289618	3.990264	8,715
1920	8,718	3.910117	−15	−59.106.25	225	886.593825	4.516314	−.225660	4.090.702	12,310
1921	19,751	4.295589	−13	−55.8.67.7	169	725.954511	4.101926	−.169196	4.232.40	17,080
1922	24,747	4.393123	−11	−48.3.7.43	121	531.601183	4.18.008	−.121.55	4.366.3	23,250
1923	32,558	4.512657	−9	−40.613911	81	365.525217	4.378410	−.081238	4.192.02	31,090
1924	12,243	4.625735	−7	−32.380435	49	226.661995	4.650772	−.049172	4.104628	10,800
1925	58,277	4.765497	−5	−23.827.85	25	119.137450	4.715701	−.025073	4.726.31	52,560
1926	60,630	4.782688	−3	−14.348064	9	43.044192	1.816736	−.009076	4.820.10	66,470
1927	100,048	5.000208	−1	−5.000208	1	5.000208	1.917.48	−.001003	4.910.45	82,520
1928	100,101	5.000438	1	5.001.18	1	5.000438	5.000.500	−.001003	5.002.97	100,600
1929	131,448	5.118753	3	15.1.2.9	9	46.068777	5.17.364	−.009026	5.1.4.91	120,300
1930	117,968	5.071766	5	25.358840	25	126.791130	5.261296	−.025073	5.21.72	111,300
1931	158,360	5.196985	7	36.3.825	49	254.617855	5.317228	−.049172	5.30.990	163,000
1932	152,041	5.181962	9	46.6.6.58	81	419.738922	5.31.215	−.081238	5.311.15	154,500
1933	211,883	5.326096	11	58.587016	121	611.157016	5.510002	−.121355	5.407.46	205,000
1934	194,771	5.289524	13	68.763812	169	893.729536	5.510002	−.169196	5.440.46	227,700
1935	252,076	5.402564	15	81.038460	225	1215.576600	5.407021	−.225660	5.1.9.4	239,500
1936	297,594	5.473624	17	93.051608	289	1581.877936	5.644955	−.289618	5.401.08	261,800
Total		87.345918		83.268812	1938	9238.383879				

Source: See Table 80.

II. $1938b = 83.268812$

          $b = .0429664$

I, $87.345918 = 18a + 1,938c.$

III $9238.383879 = 1938a + 374.034c$

(I × 107.86667) $9,404.244129 = 1938a + 208.658.0065c$

III. $9,238.383879 = 1938a + 374.034c$

       $-165.860250 = -.0010029312c$

           $c = -.0010029312$

I $87.345918 = 18a + 1,938(-.0010029342)$

        $18a = 89.289604$

          $a = 4.9605336$

Check (III): $9,238.383879 = 1,938(4.9605336) + 374.034(-.0010029342)$

                $= 9,238.3826$

Trend equation: $\log Y_c = 4.960534 + .0429664X - .0010029934X^2$

439

in absolute size.   By so doing, the trend line is more likely to go through all the cycles than merely the more recent ones.   Probably this point more than offsets the other rather technical disadvantage.

Second degree curve.   Although the rayon consumption data are concave upward when plotted on arithmetic paper, Chart 166 reveals that the use of ratio paper more than straightens out the curve, making it concave downward.   It is therefore apparent that, if logarithms of the consumption figures are used, a second degree semi-logarithmic curve in which c is negative might give a reasonably good fit.   The equation type is

$$\log Y_C = \log a + X \log b + X^2 \log c.$$

Or, for the sake of convenience, it may be written simply

$$\log Y_C = a + bX + cX^2,$$

where a, b, and c represent logarithms.

The normal equations required, when the origin is at the middle year, are

$$\text{I.} \quad \Sigma \log Y = Na + c\Sigma X^2.$$
$$\text{II.} \quad \Sigma X \log Y = b\Sigma X^2.$$
$$\text{III.} \quad \Sigma X^2 \log Y = a\Sigma X^2 + c\Sigma X^4.$$

All computations, including the Y_C values, are given in Table 98.   The method of obtaining this type of trend will be apparent from inspection of this table.   The trend values are plotted in Chart 166.   The trend line seems to fit the data reasonably well, except that it lies below the data from 1921 through 1925.   Also, this trend would eventually turn downward, which is entirely illogical.   For these reasons the second degree curve to the natural numbers is probably a better trend.

Curves with Declining Absolute Growth

It is a characteristic of many series that, although the direction of growth remains positive, the increment of growth declines with the passage of time.   A few curve types will be mentioned which fulfill the above requirement.

(1)   **Modified polynomials:** $Y_C = a + bX^{\frac{1}{2}}$.   This may be expanded as follows to include additional constants: $Y_C = a + bX^{\frac{1}{2}} + cX + \cdots$.   Of course, some of the constants may be negative, in which case the curve may ultimately turn down.

(2)   **Straight line to log X:** $Y_C = a + b \log X$.   It is difficult to find any logical justification for using logarithms of time, but occasionally such a formula gives a close fit.

(3)   **Parabolic curve to log Y:** $\log Y_C = aX^b$.   In order to fit this

curve, the formula should be stated in the following form: $\log \log Y_C = \log a + b \log X$.

(4) **Modified exponential:** $Y_C = k + ab^X$. This curve describes a series the absolute growth of which decreases by a constant proportion when a is negative and b is less than one.[4] This curve and modifications of it resulting from use of the logarithms (Gompertz curve) or reciprocals (logistic curve) of the Y values will be discussed in detail in the next section.

Asymptotic Growth Curves

Modified exponential. Not only do many series gradually taper off, but it is often true that they approach an upper limit or asymptote. Perhaps the simplest type is one in which the amount of growth declines by a

TABLE 99

HYPOTHETICAL DATA FOR MODIFIED EXPONENTIAL CURVE

(Asymptote $k = 114$)

X (1)	Y (2)	Partial totals (3)	Y increment (4)	Per cent of preceding increment (5)
0	50			.
1	66	116.0000	16	.
2	78		12	75
3	87	165.0000	9	75
4	93 75		6 75	75
5	98 8125	192.5625	5 0625	75

constant percentage. Now, the ordinary exponential (or compound interest) curve, written $Y_C = ab^X$, describes a curve the amount of change in which increases by a constant percentage if b is a positive number greater than one, but declines by a constant percentage if b is a positive number less than one. Furthermore, if the growth is declining by a constant percentage, the amount of growth approaches zero as a limit; if now we add another constant to the equation so that it reads $Y_C = k + ab^X$, the curve will approach k as a lower limit if a is positive, but k will be the upper limit if a is negative.

A series the amount of growth of which is decreasing by a constant per-

---

[4] If the absolute growth decreases by a constant proportion, then the values of the series increase at a decreasing percentage rate. Another curve which increases at a decreasing percentage rate is the type $Y_c = bc^{\frac{1}{d+X}}$. Further reference to this curve will be made in Chapter XXI.

centage is shown in the first two columns of Table 99. As can be seen in columns 4 and 5, each first difference is 75 per cent of the preceding first difference. The increments of increase are Δ_1, Δ_2, Δ_3, Δ_4, and Δ_5, and

$$\frac{\Delta_2}{\Delta_1} = \frac{\Delta_3}{\Delta_2} = \frac{\Delta_4}{\Delta_3} = \frac{\Delta_5}{\Delta_4} = .75.$$

Referring to Chart 167, the horizontal broken line near the top of the chart is the value k that the curve of this series approaches; in this case k is 114   This means that, if we should extend the trend line indefinitely
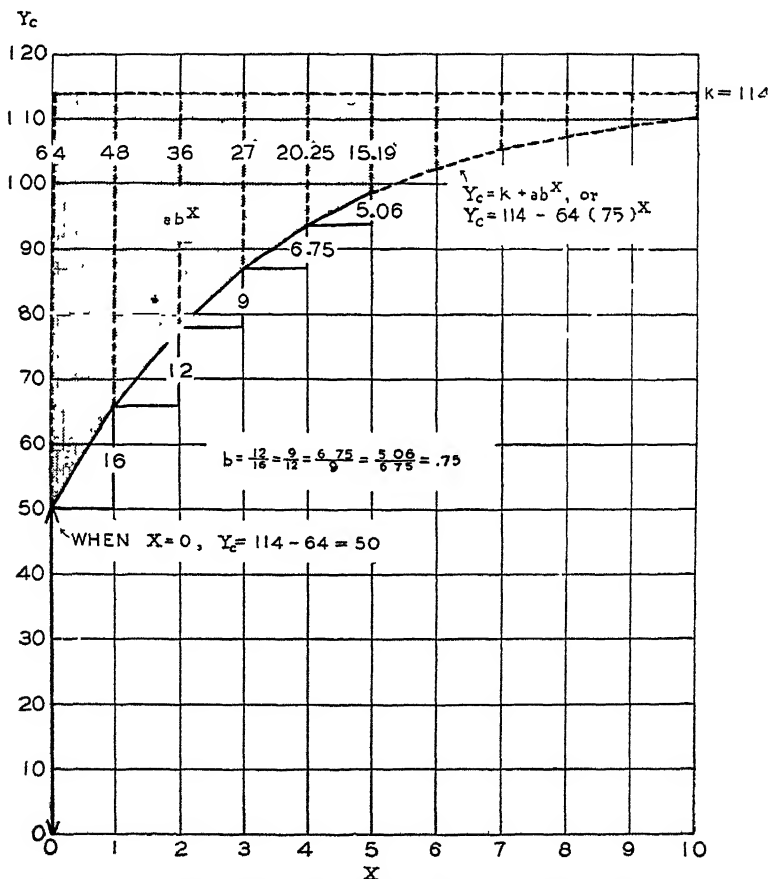
Chart 167. Artificial Data Conforming to Modified Exponential Equation.

it would approach closer and closer to this value, but never quite equal it. The second constant, a, the value obtained by subtracting the asymptote k from the trend value when X is zero, in this instance is -64. The

third constant, b, is, of course, the ratio between successive increments of growth, or .75 for this series. In Chart 167 the vertical broken line when $X = 1$ is $-64 \ (.75) = -48$; when $X = 2$ it is $-64 \ (.75)^2 = -36$; and so on for the other values of X. Thus these vertical broken lines are described by the expression ab^X. This is true when $X = 0$ also, since $-64(.75)^0 = -64$. In the diagram ab^X is represented by the height of the shaded area. If now, in turn, we subtract from k the value of each of the vertical broken lines, we have the trend values represented by large dots on the chart. The vertical broken lines are subtracted from k because the sign of a is negative. Thus:

X	$k + ab^X$	$= Y_c$
0	114 − 64	= 50
1	114 − 48	= 66
2	114 − 36	= 78
3	114 − 27	= 87
4	114 − 20 25	= 93.75
5	114 − 15 1875	= 98 8125

It is therefore evident that the equation is of the type $Y_C = k + ab^X$. The sign of a is always negative if the increments of growth are declining. As is already obvious, for this series of data the equation is $Y_C = 114 - 64(.75)^X$.

Since this curve has three constants—k, the asymptote; a, the distance between the value of Y_C when $X = 0$ and the asymptote; and b, the ratio between successive first differences—three equations are required for its fitting. These are obtained by first dividing the data into three equal sections, as in Table 99. Then the Y values are totaled for each section, as in column 3. The results are:

$$\Sigma_1 Y = 116.$$
$$\Sigma_2 Y = 165.$$
$$\Sigma_3 Y = 192.5625.$$

Let us note what 116 represents in terms of our equation. It is the sum of $50 + 66$. But 50 is $k + ab^0$ and 66 is $k + ab^1$; so

$$116 = 2k + a + ab.$$

This is equation I. The other two are obtained in similar fashion. The three equations are:

$$\text{I.} \qquad 116 = 2k + a + ab.$$
$$\text{II.} \qquad 165 = 2k + ab^2 + ab^3.$$
$$\text{III.} \quad 192.5625 = 2k + ab^4 + ab^5.$$

In order to solve for b, we first subtract equation I from equation II, ob

taining equation A; and then equation II from equation III, obtaining equation B. Thus:

$$\text{A.} \qquad 49 = ab^3 + ab^2 - ab - a$$
$$= a(b^3 + b^2 - b - 1).$$

$$\text{B.} \quad 27.5625 = ab^5 + ab^4 - ab^3 - ab^2$$
$$= ab^2(b^3 + b^2 - b - 1).$$

The constant b is now obtained by dividing equation B by equation A We shall call the resulting equation C.

$$\text{C.} \qquad \frac{27.5625}{49} = \frac{ab^2(b^3 + b^2 - b - 1)}{a(b^3 + b^2 - b - 1)}$$
$$b^2 = .5625$$
$$b = .75$$

The value of a may now be obtained by substituting in equation A or B.

$$\text{A.} \quad 49 = a(.75^3 + .75^2 - .75 - 1)$$
$$a = \frac{49}{-.765625} = -64.$$

The remaining constant k may be obtained by substituting the values of a and b in any of the original equations.

$$\text{I.} \quad 116 = 2k - 64 - 64(.75)$$
$$2k = 228$$
$$k = 114.$$

The values of the constants are thus found to be those which we knew to be correct. The equation was not obtained by the method of least squares, but was so fitted that the three partial totals of the trend values were the same as those of the original data. In this case, since the original data conform to the equation type perfectly, the fitted curve passes through all of the original points.

The logical procedure, which has been explained, can be developed into more convenient formulae, which are as follows:[5]

$$b^n = \frac{\Sigma_3 Y - \Sigma_2 Y}{\Sigma_2 Y - \Sigma_1 Y}.$$

$$a = (\Sigma_2 Y - \Sigma_1 Y)\frac{b - 1}{(b^n - 1)^2}.$$

$$k = \frac{1}{n}\left[\Sigma_1 Y - \left(\frac{b^n - 1}{b - 1}\right)a\right],$$

where n is the number of years in each third of the data.

---

[5] The derivation of these formulae is given in Appendix B. section XVI-1.

Solving by these formulae requires, of course, that b be obtained first, then a, and finally k.

The value of k may also be obtained by use of the following expressions:

$$k = \frac{1}{n}\left(\Sigma_1 Y - \frac{\Sigma_2 Y - \Sigma_1 Y}{b^n - 1}\right),$$

which does not involve the determination of a, and

$$k = \frac{1}{n}\left[\frac{(\Sigma_1 Y)(\Sigma_3 Y) - (\Sigma_2 Y)^2}{\Sigma_1 Y + \Sigma_3 Y - 2\Sigma_2 Y}\right],$$

which does not involve the determination of either a or b. These expressions for k may be derived by substituting the expressions for a and b in the equation

$$k = \frac{1}{n}\left[\Sigma_1 Y - \left(\frac{b^n - 1}{b - 1}\right)a\right].$$

Although we have used the equation $Y_C = k + ab^X$ to express the trend of a series the amount of increase in which is decreasing at a constant rate, this equation type may be used for series the amount of increase, or decrease, in which tends to decrease, or increase, at a constant rate. In any event k will always be the asymptote at $X = \pm\infty$.

Modified exponential fitted to department store sales. One of the difficulties which the world depression of the thirties occasioned for statisticians was the question of the sort of trend to use during this prolonged economic breakdown. Many trends that seemed appropriate before 1931 became rather absurd if the 1932 and 1933 data were included; for instance, this one cycle might cause the trend to turn downward. One solution that was commonly used was to refrain from revising the trend, but to extend the trend that had been fitted to data prior to 1931. Many people believed, however, that the extension of the old trend would tend to exaggerate the depression by making the trend too high. This is but another way of saying that the trend after the depression was expected to differ from the old trend, as to either its level or its slope. One conservative practice, therefore, was to select a trend type that would tend to flatten out without actually bending down. The modified exponential is well suited to do this for some series. As an illustration of method, department store sales will be used. The trend has been fitted to the 1919–1930 data and extended through 1936. The original data and the trend are plotted on Chart 168. The trend equation is

$$Y_C = 110.270 - 32.7894\,(.7814348)^X.$$

The asymptote of this curve is therefore 110.27 per cent. Although it is not likely that department store sales will always remain below this figure

TABLE 100

COMPUTATION OF MODIFIED EXPONENTIAL TREND TO FEDERAL RESERVE INDEX OF
DEPARTMENT STORE SALES

(1923–1925 = 100)

Year	X	Y (sales index)	Computation of trend values		
			b^x	ab^x	$Y_c = k + ab^x$
1919	0	78	1 0000000	−32 789	77 48
1920	1	94	7814348	−25 623	84 65
1921	2	87	6106403	−20 023	90 25
1922	3	88	4771756	−15 646	94 62
$\Sigma_1 Y$		347			347 00✓
1923	4	98	3728816✓	−12 227	98 04
1924	5	99	2913827	− 9 554	100 72
1925	6	103	2276966	− 7 466	102 80
1926	7	106	1779300	− 5 834	104 44
$\Sigma_2 Y$	.	406	.		406 00✓
1927	8	107	.1390407	− 4 559	105 71
1928	9	108	1086512	− 3 563	106 71
1929	10	111	0849038	− 2 784	107 49
1930	11	102	0663468	− 2 175	108 10
$\Sigma_3 Y$		428	.		428 01✓
1931	12	.	0518457	− 1 700	108 57*
1932	13		0405140	− 1 328	108 94*
1933	14		0316590	− 1 038	109 23*
1934	15		0247394	− 0 811	109 46*
1935	16		0193322	− 634	109 64*
1936	17	.	0151069	− 495	109 78*

* Trend values extended beyond data

Source Data for 1919–1935 from United States Department of Commerce Survey of Current Business
1936 Supplement, p 27, data for 1936 from Standard Trade and ꞏꞏꞏ ꞏ ꞏ, C ꞏ ꞏ Statistics, April 1937
p 13.

$$b^n = b^4 = \frac{\Sigma_3 Y - \Sigma_2 Y}{\Sigma_2 Y - \Sigma_1 Y} = \frac{428 - 406}{406 - 347} = \frac{22}{59} = .3728814.$$

$4 \log b = \log .3728814 = 9\ 571570 - 10 = -0\ 428430$

$\log b = -.1071075 = 9.8928925 - 10$

$b = 7814348$

$$a = (\Sigma_2 Y - \Sigma_1 Y) \frac{b - 1}{(b^n - 1)^2} = 59 \left[\frac{7814348 - 1}{(3728814 - 1)^2} \right] = 59 \left[\frac{(-2185652)}{(-6271186)^2} \right]$$

$$= -59 \left[\frac{2185652}{.3932777} \right] = -59(5557528) = -32\ 7894152.$$

$$k = \frac{1}{n} \left[\Sigma_1 Y - \left(\frac{b^n - 1}{b - 1} \right) a \right] = \frac{1}{4} \left[347 - \frac{-6271186}{-.2185652} (-32\ 7894152) \right]$$

$$= \frac{1}{4} [347 - (2\ 8692518)(-32.789415)]$$

$$= \frac{1}{4} (347 + 94.0810880) = \frac{441.08108}{4} = 110.270.$$

$Y_c = 110\ 270 - 32.7894152\ (.7814348)^x,$

with origin at 1919 and X units 1 year.

nevertheless it is not an unreasonable equation to use until business regains sufficient stability to permit a more appropriate trend to be selected.

The computation of this trend is given in Table 100. The constants are obtained by means of the formulae on page 444. The last three columns of the table are for obtaining the trend values. Column b^x is obtained by placing the value of b in the keyboard of the calculating machine and multiplying it by the value of b^x last obtained. It is advisable to obtain column b^x as soon as the value of b is obtained. If no mistakes have been made, the value of b^n in the table will agree with that obtained

PER CENT
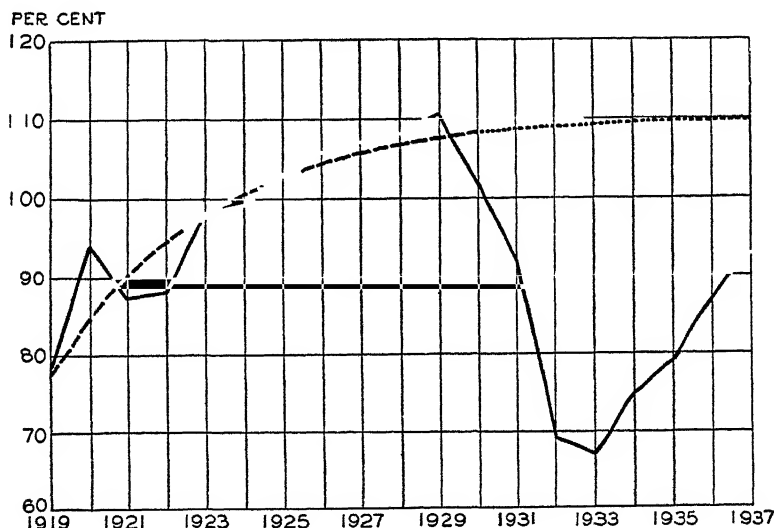
Chart 168.   Modified Exponential Trend Fitted to Federal Reserve Index of Department Store Sales, 1919–1930, and Trend Extension Through 1937.   (Data of Table 100 and United States Department of Commerce, *Survey of Current Business*, March 1938, p. 27.)

by the formula. Column ab^x is obtained by putting the value of a in the keyboard and multiplying by the appropriate b^x values. A final check on the accuracy of the work is obtained by summing each of the three sets of Y_C values. (Note check marks in Table 100.) This is equivalent to verifying by substituting the value of the constants in the normal equations.

Gompertz curve. More commonly used than the modified exponential just described is the Gompertz curve. While the modified exponential has to do with a series the growth increments of which are declining by a constant percentage, the Gompertz curve describes a series in which the growth increments of the logarithms are declining by a constant percentage

The natural values of the series described by the Gompertz curve show a declining rate of growth, but the rate does not decline by either a constant amount or a constant percentage. A second degree curve may also describe a series whose rate of increase is decreasing, but such a curve does not flatten out at the top. Furthermore, while the second degree curve is not asymptotic, and the modified exponential has only an upper limit, the Gompertz curve is asymptotic at both ends, the lower asymptote being zero.

A Gompertz curve fitted to the rayon data from 1919 through 1936 and extended in both directions to include 1915 and 1950 is shown in Chart 169. The extensions are for the purpose of illustrating the shape of the curve; no prediction is intended. It will be noticed that the amount of growth is small at first, then becomes larger until it reaches a point of inflection, after which it declines and finally approaches, but never reaches, zero. This general shape of the trend is common to many industries and has led Prescott[6] to the conclusion that it describes a law of growth. According to Prescott, this trend is a function of population growth, the curve of which typically is similar in appearance, but it is also partly due to the development of the individual industry. He believes that the growth of an industry may be divided into four stages:

(1) Period of experimentation.
(2) Period of growth into the social fabric.
(3) Through the point where growth increases but at a diminishing rate.
(4) Period of stability.

Although these stages are not very specifically demarcated by Prescott, apparently rayon consumption is now in the third stage, for it is not until 1935 that the increment of trend growth begins to fall off. Prescott also claims for this type of curve that it is useful in forecasting the future of an industry, since it is not only a logical curve but, on account of its tendency to flatten out, it tends to be conservative in its forecasts. The horizontal dashed line of section A of Chart 169 indicates that the upper limit of United States rayon consumption will eventually be about 520,000,000 pounds per year. It seems quite likely, however, that the Gompertz curve is unduly conservative in this instance.

The same data and trend are shown on ratio paper in section B of Chart 169. On this type of paper the resemblance to the modified exponential curve (drawn on arithmetic paper) is apparent. In fact the Gompertz curve is exactly like the modified exponential except that it is the increments of increase in the logarithms of the Y values that are declining at

---

[6] "Law of Growth in Forecasting Demand," by Raymond D. Prescott. *Journal of the American Statistical Association*, Vol. XVIII, December 1922, pp. 471–479.
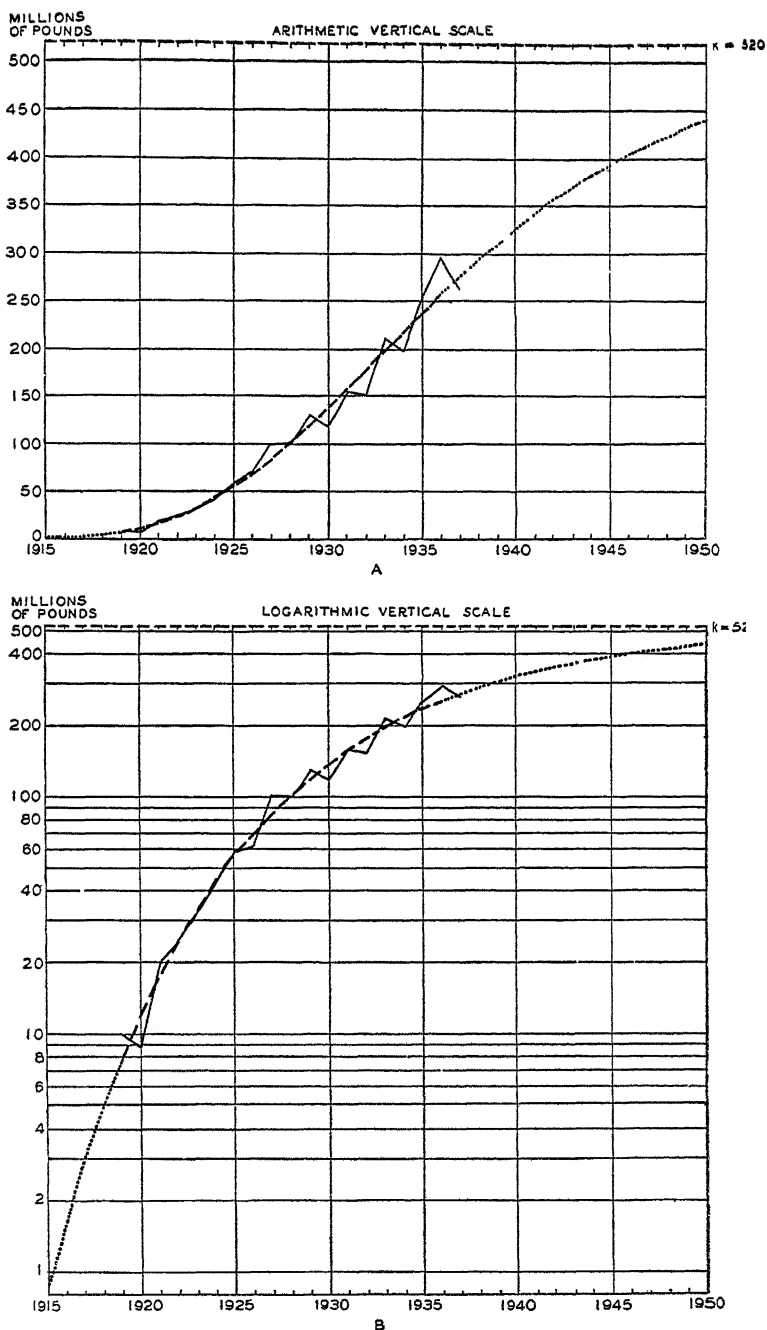
Chart 169. Gompertz Curve Fitted to United States Rayon Consumption, 1919-1936:
A. Arithmetic Vertical Scale; B. Logarithmic Vertical Scale. (Trend is extended to
1915 and to 1950 to show general shape of curve.   Data of Tables 80 and 101.)

a constant rate. Consequently the equation type[7] may be written

$$\log Y_C = \log k + (\log a)b^X;$$

and the constants are obtained by the following formulae:

$$b^n = \frac{\Sigma_3 \log Y - \Sigma_2 \log Y}{\Sigma_2 \log Y - \Sigma_1 \log Y}.$$

$$\log a = (\Sigma_2 \log Y - \Sigma_1 \log Y)\frac{b-1}{(b^n-1)^2}.$$

$$\log k = \frac{1}{n}\left[\Sigma_1 \log Y - \left(\frac{b^n-1}{b-1}\right)\log a\right],$$

$$= \frac{1}{n}\left(\Sigma_1 \log Y - \frac{\Sigma_2 \log Y - \Sigma_1 \log Y}{b^n - 1}\right),$$

$$= \frac{1}{n}\left[\frac{(\Sigma_1 \log Y)(\Sigma_3 \log Y) - (\Sigma_2 \log Y)^2}{\Sigma_1 \log Y + \Sigma_3 \log Y - 2\Sigma_2 \log Y}\right].$$

Since these formulae exactly parallel those of the modified exponential curve, no explanation of their use is necessary. However, if the reader will refer to Table 101, he will find there all of the important computations necessary to obtain the trend equation for rayon consumption. This equation is

$$\log Y_C = 5.716174 - 1.8259494 \, (.9002623)^X$$

with origin at 1919 and X units of one year.

This equation type may also be written in the form:

$$Y_C = ka^{b^X}.$$

In order to put the rayon equation in this form, it is necessary to look up the anti-logs of the constants k and a, which are in logarithmic form. Since 5.716174 is the log of 520,205 and -1.8259494, or $8.1740506 - 10$, is the log of .0149297, the equation becomes

$$Y_C = 520,205(.0149297)^{.9002623^X}.$$

Since b is .9002623, increments of growth in the logarithms of the trend values are each 90.02623 per cent of the preceding year. The value of b will always be less than one if the rate of growth of the series is declining. Since $b - 1$ will be negative under those circumstances, so will log a (see equation for log a); and a will be less than 1. Therefore the greater the value of X, the smaller becomes the value of b^X. As this value approaches

---

[7] While generally used for data the logarithms of which tend to increase by amounts which are decreasing at a constant rate, it may be applied to data the logarithms of which tend to increase, or decrease, by amounts which are decreasing, or increasing, at a constant rate

TABLE 101

COMPUTATION OF GOMPERTZ EQUATION FITTED TO UNITED STATES CONSUMPTION OF
RAYON, 1919–1936

(Thousands of pounds)

Year	X	Con-sumption Y	log Y	Computation of trend values			
				b^X	$(\log a)\, b^X$	$\log Y_c = \log k + (\log a)\, b^X$	Y_c
1919	0	9,291	3 968062	1 0000000	−1 825949	3 890225	7,766
1920	1	8,718	3 940417	9002623	−1.643833	4 072341	11,810
1921	2	19,751	4 295589	8104722	−1 479881	4 236293	17,230
1922	3	24,747	4 393423	7296376	−1.332281	4.383893	24,200
1923	4	32,558	4 512657	6568652	−1.199403	4 516771	32,870
1924	5	42,243	4 625755	5913510	−1 079777	4.636397	43,290
Σ_1 log Y			25.735903	.		25 735920 √	
1925	6	58,277	4 765497	5323710 √	− 972083	4.744091	55,470
1926	7	60,630	4 782688	4792735	− .875129	4 841045	69,350
1927	8	100,048	5 000208	4314719	− 787846	4 928328	84,790
1928	9	100,101	5,000438	3884379	− 709268	5 006906	101,600
1929	10	131,448	5 118753	3496960	− 638527	5.077647	119,600
1930	11	117,968	5 071766	3148181	− 574842	5 141332	138,500
Σ_2 log Y			29 739350			29.739349 √	
1931	12	157,360	5 196895	2834189	− 517509	5.198665	158,000
1932	13	152,041	5 181962	2551514	− 465894	5 250280	177,900
1933	14	211,883	5 326096	2297032	− 419426	5 296748	198,000
1934	15	197,771	5.289524	.2067931	− .377594	5 338580	218,100
1935	16	252,676	5.402564	1861680	− .339933	5 376241	237,800
1936	17	297,594	5 473624	1676000	− .306029	5 410145	257,100
Σ_3 log Y			31 870665			31 870659 √	

Source See Table 80.

$$b^n = \frac{\Sigma_3 \log Y - \Sigma_2 \log Y}{\Sigma_2 \log Y - \Sigma_1 \log Y}.$$

$$b^6 = \frac{31\ 870665 - 29.739350}{29.739350 - 25\ 735903} = \frac{2\ 131315}{4\ 003447} = 53236998.$$

$$\log b^6 = 9\ 72621338 - 10 = -.27378662$$

$$\log b = -\ 045631103 = 9.954368897 - 10$$

$$b = .9002623$$

$$\log a = (\Sigma_2 \log Y - \Sigma_1 \log Y)\ \frac{b - 1}{(b^n - 1)^2}$$

$$= 4.003447\ \frac{.9002623 - 1}{(.53236998 - 1)^2} = 4\ 003447\ \frac{-\ 0997377}{(-\ 46763002)^2}$$

$$= -1.8259494,\quad \text{or } \bar{2}.1740506,\quad \text{or } 8.1740506 - 10.$$

$$\log k = \frac{1}{n}\left[\Sigma_1 \log Y - \left(\frac{b^n - 1}{b - 1}\right)\log a\right]$$

$$= \frac{1}{6}\left[25.735903 - \frac{-.46763002}{-\ 0997377}\ (-1.8259494)\right]$$

$$= 5\ 716174.$$

Trend equation:

$$\log Y_c = 5\ 716174 - 1.8259494(.9002623)^X$$

$$Y_c = 520,205(.0149297)^{.9002623^X},$$

with origin at 1919 and X units 1 year.

zero, the value of a^{b^X} approaches 1, and Y_C approaches k, or 520,205, which is the upper asymptote. On the other hand, when X is zero, a^{b^X} has the same value as a, which is .149297, and Y_C is 520,205 $\times$ .149297 = 7,766.5.

A FIRST DIFFERENCES OF A GOMPERTZ CURVE

B FIRST DIFFERENCES OF A LOGISTIC CURVE
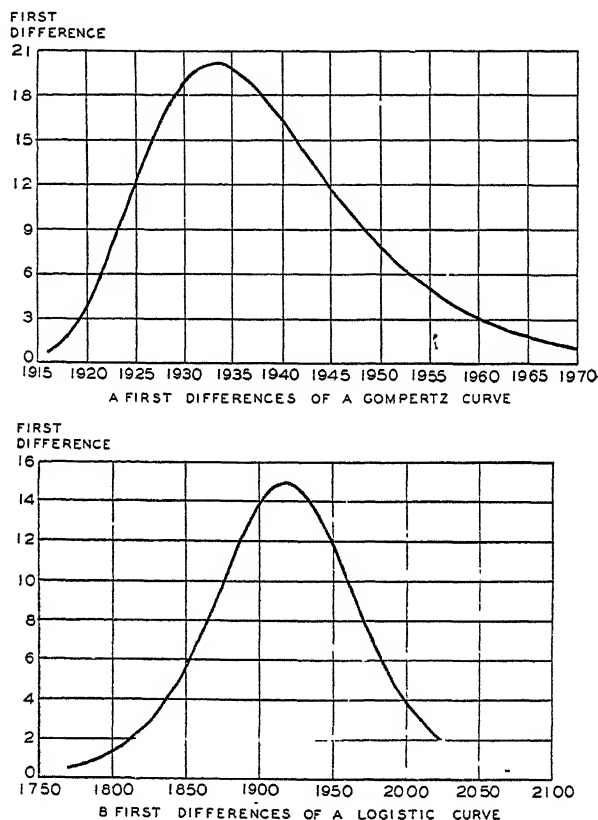
Chart 170. Growth Increments of a Gompertz Curve and of a Logistic Curve. (Gompertz curve increments are first differences of trend values of Table 101, while Logistic Curve Increments are first differences of trend values of Table 102. Trend values have been extrapolated to show better the shapes of the curves. In each case vertical distances represent changes from preceding year.)

Logistic curve. Another type of growth curve which has the same general shape as the Gompertz curve is the logistic. It is, in fact, identical with the modified exponential except that it is the first differences of the reciprocals which are declining by a constant percentage. A modified exponential may therefore be fitted by the method of partial totals to the reciprocals of the Y values, and the reciprocals of the fitted values so

obtained may be taken as the trend values.   The equation type may therefore be

$$\frac{1}{Y_C} = k + ab^X.$$

More often, however, this curve is fitted by the method of selected points. Such a process is more subjective, but is probably more generally used. When fitted by this method, the equation type is

$$Y_C = \frac{k}{1 + e^{a + bX}}.$$

In this equation k, which is one of the constants to be discovered, is the upper limit, and e is 2.71828, the base of the Naperian system of logarithms. Since b is negative, the value of $a + bX$ must eventually become negative and $e^{a + bX}$ a very small fraction, so that Y_C approaches the value k (is asymptotic to k).   Consequently k is the upper limit.   On the other hand, for large negative values of X, the denominator will become very large and the trend value approach zero.

Although the logistic is like the Gompertz in important respects, there is one dissimilarity that is easy to observe.   The first differences of the logistic curve when plotted produce a symmetrical curve that closely resembles a normal frequency distribution, while those of the Gompertz curve are skewed.   Chart 170, on which are plotted (in part A) the first differences of the rayon data Gompertz trend and (in part B) those of the logistic trend of United States population, brings out this point very clearly.

The logistic curve is fitted by a method that makes the curve pass through three subjectively selected points equidistant from each other: one near the beginning of the period, one in the middle, and one near the end.   The computation of the trend values for United States population growth is shown in Table 102.   The three selected years are 1800 (x_0), 1860 (x_1), and 1920 (x_2).   The y values chosen are the geometric three-decade averages centering on these periods.   Averages were used in order to eliminate abnormal values.   The geometric mean was used in preference to the arithmetic mean since the growth is more nearly straight line geometrically than arithmetically.   There is no certainty, however, that this method is an improvement over the selection of values entirely from inspection of the chart.   The averages obtained are as follows:

$$y_0 = \quad 5,325.$$
$$y_1 = \quad 30,408.$$
$$y_2 = 106,079.$$

Designating by n the number of years from x_0 to x_1, or from x_1 to x_2, the

TABLE 102

Logistic Curve Fitted to United States Population, 1790–1930

Year (1)	Selected point of time x (2)	X (3)	Population in thousands Y (4)	Selected y value* (5)	Computation of trend values				
					.1366265X (6)	$\log \mu =$ 1 542035 − .1366265X (7)	μ (8)	$1 + \mu$ (9)	$Y_c =$ $\dfrac{190{,}830\ 35}{1 + \mu}$ (10)
1780		−2			− 273253	1 815288	65 356	66 356	2,876
1790		−1	3,929		− 136626	1 678662	47 716	48 716	3,917
1800	x_0	0	5,308	5,325 (y_0)	0	1.542035	34 837	35.837	5,325 √
1810		1	7,240		136626	1 405408	25 433	26 433	7,220
1820		2	9,638		273253	1 268782	18 564	19 564	9,754
1830		3	12,866		409880	1 132156	13.557	14 557	13,109
1840		4	17,069		546506	995529	9 8976	10.8976	17,511
1850		5	23,192		683132	858902	7 2261	8 2261	23,198
1860	x_1	6	31,443	30,408 (y_1)	819759	722276	5 2757	6 2757	30,408 √
1870		7	38,558		956386	585650	3 8517	4.8517	39,333
1880		8	50,156		1 093012	449023	2 8121	3 8121	50,059
1890		9	62,948		1 229638	312396	2.0530	3 0530	62,506
1900		10	75,995		1 366265	175770	1 4989	2 4989	76,366
1910		11	91,972		1 502892	039144	1 0943	2 0943	91,119
1920	x_2	12	105,711	106,079 (y_2)	1 639518	−.097483	.79895	1 79895	106,079 √
1930		13	122,775		1 776144	− 234110	58330	1 58330	120,527
1940		14			1 912771	− 370736	42586	1 42586	133,835
1950		15			2 049398	− 507362	31091	1 31091	145,571

* Each y value is the geometric mean of three values in column 4 Thus:

$y_0 = (3{,}929 \times 5{,}308 \times 7{,}240)^{\frac{1}{3}} = 5{,}325.$

$y_1 = (23{,}192 \times 31{,}443 \times 38{,}558)^{\frac{1}{3}} = 30{,}408$

$y_2 = (91{,}972 \times 105{,}711 \times 122{,}775)^{\frac{1}{3}} = 106{,}079.$

formulae required for computing the constants for the logistic curve are:[8]

$$k = \frac{2y_0y_1y_2 - y_1^2(y_0 + y_2)}{y_0y_2 - y_1^2}.$$

$$a = \log_e \frac{k - y_0}{y_0}.$$

$$b = \frac{1}{n} \log_e \frac{y_0(k - y_1)}{y_1(k - y_0)}.$$

Substituting in the first equation, we have:

$$k = \frac{2(5,325)(30,408)(106,079) - (30,408)^2(5,325 + 106,079)}{5,325(106,079) - (30,408)^2}$$

$$= 190,830.35.$$

The second equation becomes

$$a = \log_e \frac{190,830.35 - 5,325}{5,325} = \log_e \frac{185,505.35}{5,325} = \log_e 34.836685.$$

However,

$$\log_e X = 2.302585 \log_{10} X.$$

Therefore

$$\log_e 34.836685 = 2.302585 \log_{10} 34.836685,$$

and

$$a = 2.302585 \log 34.836685 = 3.5506713.$$

Finally, substituting in the last equation

$$b = \frac{1}{6} \log_e \frac{5,325(190,830.35 - 30,408)}{30,408(185,505.349)} = \frac{1}{6} \log_e .15143986$$

$$= -.3145945.$$

The trend equation therefore may be stated

$$Y_C = \frac{190,830.35}{1 + e^{3.550671 - .3145945X}},$$

with origin at 1800 and X units 10 years.

In obtaining the trend values, it is possible to save one column of multiplications by simplifying the formula. If we designate by μ the expression $e^{a + bX}$, the formula becomes

$$Y_C = \frac{k}{1 + \mu}.$$

---

[8] For the mathematical reasoning behind this type of curve, see Raymond Pearl, *Studies in Human Biology*, Chapter XXIV. Williams and Wilkins Company, Baltimore, 1924.

In our equation

$$\mu = e^{3.550671 - .3145945X},$$

and

$$\log_{10} \mu = (3.550671 - .3145945X) \log_{10} e$$
$$= .43429(3.550671 - .3145945X)$$
$$= 1.542035 - .1366265X.$$

In Table 102 the values for μ are computed in columns 6, 7, and 8. A final check on the computations may be had by comparing the Y_C values for x_0, x_1, and x_2 with the values of y_1, y_2, y_3, since the curve must pass through the three selected points. The check marks in column 10 of the table indicate perfect agreement.

The results of the curve fitting are shown in Chart 171. Since the method of fitting is based on selected points, the fitted curve would of necessity coincide with the values selected for those points. The chart, however, shows extremely close relationship throughout. The trend has been extended a number of decades in order to show more completely the fundamental shape of the curve.

The logistic curve owes its name to a Belgian mathematician by the name of Verhulst, who used it as early as 1838 as an expression of the law of population growth, and gave it that name. In recent years it has been used extensively by the biometricians Raymond Pearl and L. J. Reed, and is frequently called the *Pearl-Reed curve*. They have used it to describe the growth of an albino rat, a tadpole's tail, the number of yeast cells in a nutritive solution, the number of fruit flies in a bottle (on a limited food supply), and, most interesting of all, the number of human beings in a geographical area. In each case the phenomenon measured is population growth, either the number of cells in an organism or the number of individuals in a region. The law of growth which the logistic curve describes is stated by Pearl as follows:[9]

> In a spatially limited universe the amount of increase which occurs in any particular unit of time, at any point of the single cycle of growth, is proportional to two things, viz: (a) the absolute size already attained at the beginning of the unit interval under consideration, and (b) the amount still unused or unexpended in the given universe (or area) of actual and potential resources for the support of growth.

In the case of human populations some development may expand the available subsistence and allow a new cycle of growth. For instance, mankind may pass through a hunting stage, an agricultural stage, and

---

[9] Raymond Pearl, *The Biology of Population Growth*, Alfred A. Knopf. New York, 1925, p. 22.

an industrial stage.   Each cultural epoch may then be described by a new logistic curve spliced onto the old one.   Thus

$$Y_C = k_1 + \frac{k_2}{1 + e^{a + bX}}$$
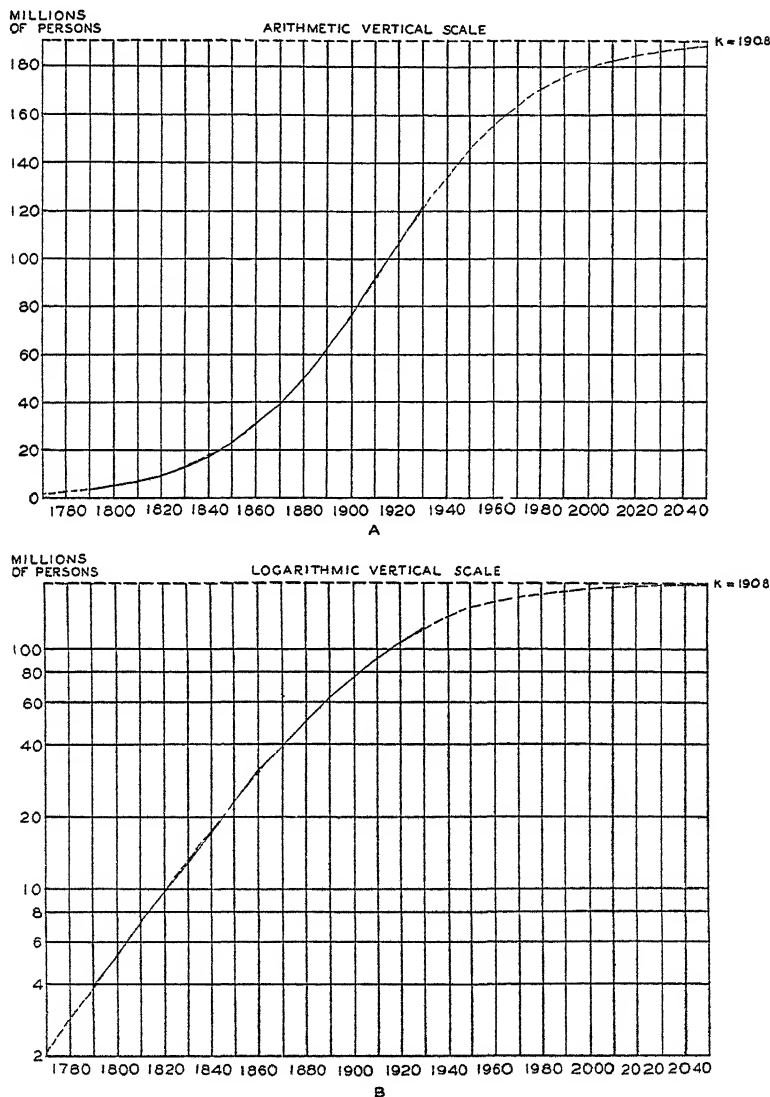
Chart 171.   Logistic Trend Fitted to Population Growth, Continental United States. 1790–1930.   (Dotted lines 1770–1790 and 1930–2050 indicate extensions beyond data to show general shape of curve.   Data of Table 102.)

describes a curve in which k_1 is the new lower limit and $k_1 + k_2$ the new upper limit. In this equation, k_1 is below the upper limit k_0 of the previous logistic and indicates the value at which the previous one was interrupted.

Apparently waves of immigration and human institutions do not change the fundamental shape of the curve, although the steepness of its slope may be modified somewhat. Also the growth may not be symmetrical: the point of inflection need not be halfway between the upper and the lower asymptotes, nor need the two parts of the curve be of the same shape. A skewed logistic may be obtained by a slight modification of the previous formulae:

$$Y_c = k_1 + \frac{k_2}{1 + e^{a\,+\,bX\,+\,cX^2}}.$$

The theory advanced by Raymond Pearl is not, however, universally accepted. Some argue that, although the logistic curve is appropriate enough for fruit flies in a bottle, its extension to human society is unwarranted. Human beings have, and exercise, the power of modifying their environment and rationally controlling their rate of reproduction.

One of the chief objects of the logistic curve is the forecasting of future growth. Thus Table 102 shows the trend values extended through 1950. But it may also be used to estimate the population for earlier periods, before the existence of reliable records. Thus the population of the region later to become the United States is estimated in Table 102 to have been 2,876,000 in 1780. Of course, extensions of the curve give reliable results only if there are no changes in the area involved and no new influences arise to affect the rate of population growth; and the accuracy varies inversely with the extent of extrapolation.

Use of arithmetic probability paper. Attention has already been called to the close resemblance between the first differences of the curves described in this section and the ordinary frequency polygon. Chart 170 could easily be mistaken for two such curves. No doubt also the reader has noticed that the trend values when plotted on ordinary arithmetic paper are very similar to ogives, or cumulative frequency polygons. Since this is true, it might well be that some series, if plotted on arithmetic probability paper, would approximate a straight line. Arithmetic probability paper, however, is in terms of percentages, and either the scale of the probability paper must be converted into the units of the data being used, or the data must be converted into percentages. The latter is easier. Since 100 per cent is the theoretical upper limit of probability paper, it is necessary only to assume some upper limit for the time series in question and to express the value for each year as a percentage of this maximum.

Using the population data as an illustration, let us assume that 150,000,000 is the maximum figure for our population. Dividing this number by each of our decennial population counts gives the percentages of Table 103, column 3. These percentages are plotted on Chart 172. The result is a curve that is concave upward. The curve can be straightened out, however, by taking a larger upper limit. Thus upper limits, in turn,
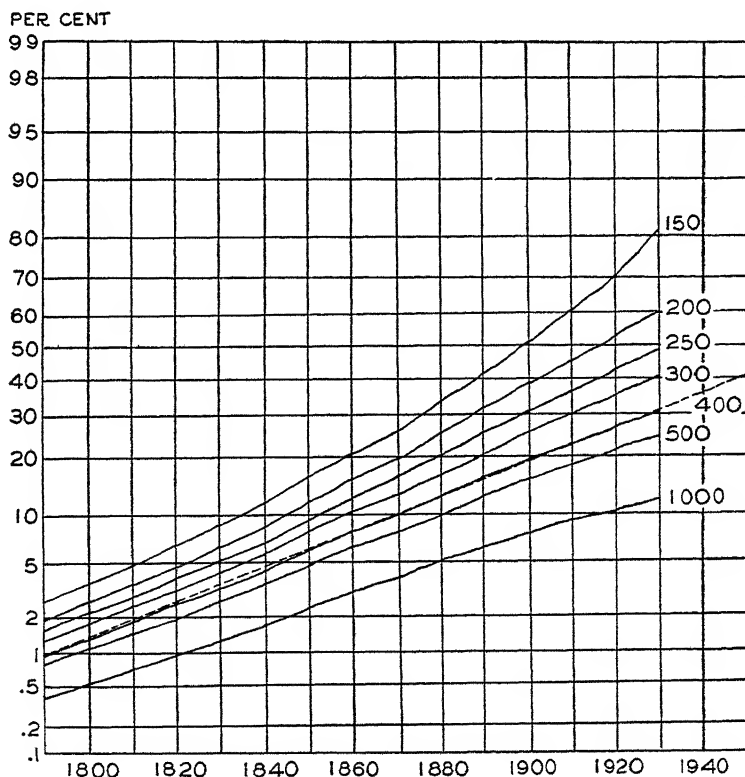
Chart 172. Estimating the Trend of United States Population by Use of Arithmetic Probability Paper. (Trial values of upper limit are shown to the right in the chart. Data of Table 103.)

of 150, 200, 250, 300, 400, 500, and 1,000 millions (shown to the right in this chart) produce the different curves of Chart 172. The curves gradually becomes straighter as the one with 400 millions as a limit is approached, after which they begin to be slightly concave downward. Since the 400 line seems to be most nearly straight, a straight line is drawn through the points on this line and, in order to estimate future population, the line is extended through 1950. Readings from this line are recorded

in column 5 of Table 103.   The trend values are now obtained simply by multiplying these decimals by 400,000,000.

Although the trend values are not greatly d'fferent from those obtaine d by the logistic curve until the curve is extended far beyond the data, the upper limit of population growth by the graphic method is more than double that obtained by the mathematical equation.   Most experts on population growth would probably say that 190,000,000 is closer to the truth than 400,000,000.

TABLE 103

ESTIMATING THE TREND OF UNITED STATES POPULATION FROM PROBABILITY PAPER

Year (1)	Population (2)	Population as per cent of 150,000,000 (3)	Population as per cent of 400,000,000 (4)	Trend readings from Chart 172 (per cent) (5)	Y_c [400,000,000 × column 5] (6)
1790	3,929,000	2 61	0 98	.90	3,600,000
1800	5,308,000	3.54	1 33	1 35	5,400,000
1810	7,240,000	4.83	1 81	1 80	7,200,000
1820	9,638,000	6 43	2 41	2 50	10,000,000
1830	12,866,000	8.58	3 22	3 35	13,400,000
1840	17,069,000	11.38	4.27	4.55	18,200,000
1850	23,192,000	15.46	5 80	5 95	23,800,000
1860	31,443,000	20 96	7 86	7.70	30,800,000
1870	38,558,000	25.71	9 64	9.80	39,200,000
1880	50,156,000	33.44	12 54	12.45	49,800,000
1890	62,948,000	41 97	15 74	15.35	61,400,000
1900	75,995,000	50 66	19 00	18 80	75,200,000
1910	91,972,000	61 31	22 99	22 40	89,600,000
1920	105,711,000	70 47	26.43	27.00	108,000,000
1930	122,775,000	81.85	30.69	31 40	125,600,000
1940	...	..	...	36.15	144,600,000
1950	...	..	..	41 50	166,000,000

Source: See Table 102

Thus Pascal K. Whelpton, a member of the American delegation at the Population Congress held in Paris in July 1937, is quoted as saying in an interview:[10]

United States population experts seem to agree that the United States will reach the climax of its population in the next generation.   Perhaps the top figure will be 150,000,000.   After that the population is bound to decrease.   That it is already shrinking is evidenced by the fact that there are fewer children today in the first five grades of the public schools.

[10] As reported by the New York *Times*, July 31. 1937.

The United States population has increased only 9,000,000 from 1930, owing to the practice of birth control and lessened immigration. Most of us agree it is a good thing, leading to a higher standard of living.

Expert opinion also leans toward methods of population prediction which put less reliance on extending curves and more on analysis of the causal factors where these are known. The chief of these factors are:

1. *Births.* The number of births in a country bears a relationship, not to the number of persons in the country, but to the number of women of child-bearing age. Thus the age and sex distribution of a country need be considered, as well as any trend in this ratio.

2. *Deaths.*

3 *Immigration.*

4. *Emigration.*

By considering such factors as these, including their estimated trends in the future, most statisticians obtain a maximum well under 200,000,000. It is also considered possible that the population of the United States may decrease before the year 2,000, since the women of child-bearing age are not producing enough offspring to maintain themselves.

The factors mentioned above can be utilized for the United States as a whole, but the last two factors, immigration and emigration, are not available for individual states and cities. In these uses, interpolation and extrapolation must be made by means of a trend equation or by reading from a curve.

Objective Tests of Trends

It must not be imagined that this chapter is at all exhaustive of the types of trends that may be utilized. However, a sufficient variety has been given to meet most of the needs for time series analysis. Since such a large number are available, how can we decide which to use? First of all, let it be repeated that we should select a trend which describes the forces sought to be measured. If the object is solely to obtain cyclical deviations, probably the trend should pass through the approximate center of each cycle. If the object is forecasting, a mathematical equation should be selected which, when extended, will conform to expectations dictated by logic. If, for instance, the series is such that it may logically be expected to flatten out, an asymptotic curve should be selected. If the object is solely historical study, the future behavior of the curve is not so important.

Assuming that economic processes conform to some "law," this law may perhaps be discovered by first smoothing the data somewhat and then applying certain objective tests:

(1) If the first differences are constant, use a straight line.

(2) If the second differences are constant, use a second degree curve.

(3) If the third differences are constant, use a third degree curve.

(4) If the first differences are changing by a constant percentage, use a modified exponential.

(5) If the first differences resemble a normal curve, use a logistic.

(6) If the first differences resemble a skewed frequency curve, use a Gompertz curve or a complex type of logistic.

(7) If the first differences of the logarithms are constant, use an exponential. (Fit a straight line to the logarithms.)

(8) If the second differences of the logarithms are constant, fit a second degree curve to the logarithms.

(9) If the first differences of the logarithms are changing by a constant percentage, use a Gompertz curve.

(10) If the first differences of the reciprocals are changing by a constant percentage, use a logistic curve.

It is not usually necessary to make these tests. Plotting the original data on arithmetic paper, semi-logarithmic paper, or probability paper may dictate the choice. It may be, however, that neither inspection of such charts nor the more objective tests we have described will be conclusive. In the first place, the preliminary smoothing may not have been done properly. Secondly, the series may not conform to any simple mathematical description. In a dynamic world, forces in operation are seldom allowed to work themselves out before other factors make themselves felt. Consequently any type of mathematical trend may "work" only for a relatively short period.

In case of doubt as to which of several trends (each of the same number of constants) to use, that one is to be preferred from which the sum of the squared deviations of the data is at a minimum. In making this comparison, arithmetic curves should not be compared with those fitted to logarithms.

Selected References

H. G. Brunsman: *Simplified Procedure in the Statistical Analysis of Time Series;* Ohio State University Press, Columbus, Ohio, 1930. Discusses summation method.

G. R. Davies and W. F. Crowder: *Methods of Statistical Analysis in the Social Sciences,* Chapter VI; John Wiley and Sons, New York, 1933. Includes summation method of fitting polynomials.

R. A. Fisher: *Statistical Methods for Research Workers* (Seventh Edition), pp. 148–158; Oliver and Boyd, Edinburgh, 1936. A discussion of orthogonal polynomials and the fitting of polynomials by a summation method.

Simon Kuznets: *Secular Movements in Production and Prices;* Houghton Mifflin Co., Boston, 1930. The concepts of primary and secondary trends are explained and illustrated.

F. C. Mills: *Statistical Methods Applied to Economics and Business* (Revised Edition), pages 253–279 and Appendix D; Henry Holt and Co., New York, 1938.   Contains a section on selection of curve type.   Appendix D illustrates another use of the modified exponential, and use of reciprocals in fitting a logistic.

Raymond Pearl· *Studies in Human Biology*, Chapters XXIV, XXV; Williams and Wilkins, Baltimore, 1924.   Chapter XXIV explains the theory of the logistic curve, while Chapter XXV shows its application to the growth of human populations.

C. H. Richardson· *An Introduction to Statistical Analysis*, pages 169–200; Harcourt, Brace and Co., New York, 1934.   Gives properties and methods of fitting a number of types of curves.

T. R. Running. *Empirical Formulas;* John Wiley and Sons, New York, 1917.   The properties of a considerable number of curves are given, as well as directions for fitting them.   Graphic devices are frequently employed.

Max Sasuly: *Trend Analysis of Statistics;* Brookings Institution, Washington, 1934.   An advanced treatise for students trained in mathematics. Some of the material has to do with cyclical movements.

J. G. Smith· *Elementary Statistics*, Chapters VI, X, XII, XIV; Henry Holt and Co., New York, 1934.   Chapter VI is an elementary treatment of the mathematical meaning of certain types of equations.   Chapter X gives a brief treatment of the method of least squares and of curve fitting.   In Chapter XII, empirical trends, their validity, and methods of fitting are discussed.   Chapter XIV, pp. 260–262, explains how to convert second degree trends from an annual to a monthly basis.

G. W. Snedecor: *Statistical Methods Applied to Experiments in Agriculture and Biology*, pages 279-289; Collegiate Press, Ames, 1937.   Gives directions for fitting orthogonal polynomials as high as seventh degree.

Carl Snyder *Business Cycles and Business Measurements*, Chapter II; Macmillan Co., New York, 1927.   Shows the results of fitting trends to a variety of data.

F. F. Stephen: "Summation Methods in Fitting Parabolic Curves," *Journal of the American Statistical Association*, Vol. XXVIII, December 1932, pp. 413-423.

CHAPTER XVII

PERIODIC MOVEMENTS

As indicated in Chapter XIV, there are many types of periodic movements, including those that repeat themselves daily, weekly, monthly, or annually. In this chapter most of the measurements will be of monthly movements within a year, commonly known as seasonal movements. The principles laid down can easily be applied to the various other types of periodic movements. It will be the plan of this discussion to start with data which lend themselves to very simple treatment, and gradually to introduce more complex methods as they are required. The treatment of seasonal movements that vary in their pattern from year to year will, however, be reserved for the following chapter, as will the measurement of weekly seasonal movements which involves special problems. In general, all the methods involve averaging separately the values of the different Januaries, Februaries, etc., but differ chiefly in the degree to which the data are refined before being averaged.

Averages of unadjusted data. When the data do not contain cyclical movements or trend to any appreciable extent, it will suffice to average the data without making any previous adjustment. An illustration of such data is the circulation of books in the reserve room of the University of North Carolina, which is shown in part A of Table 104. From this table were excluded data for part weeks and the week preceding spring vacation—the latter because circulation was unusually low during the last two days of school. Below each column of data is given the average of that column. The averages, one for each day of the week, constitute a measure of the intra-week fluctuation in circulation of books. For convenience, however, it may be desirable to express this measure in percentage form. By dividing each of the seven daily averages by the average weekly circulation and multiplying by 100, we find the per cent of total weekly circulation for each day of the week. Thus, only 5.4 per cent of the total normally occurs on Sunday, while on Monday it averages 19.8 per cent. A more usual method of stating the index is to express each

TABLE 104

COMPUTATION OF INDEXES OF INTRA-WEEK VARIATION OF CIRCULATION OF BOOKS AT THE UNIVERSITY OF NORTH CAROLINA RESERVE ROOM, SPRING QUARTER, 1937

A Unadjusted data (number of books)

Week ended: (or description of row)	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Total	Average
1. March 27	214	657	643	635	667	331	249	3389	484 14
2. April 3	268	649	490	444	486	296	302	2935	419 29
3. May 1	160	714	610	550	474	160	333	3301	471 57
4. May 8	153	530	590	474	387	225	206	2565	366 43
5. May 15	137	642	626	632	639	333	317	3326	475 14
6. May 22	139	826	778	589	690	461	366	3852	550 29
7. May 29	165	538	664	546	697	278	551	3639	519 86
8. Arithmetic mean	176 57	650 86	628 71	552 86	577 14	370 00	330 57	3286 71	
9. Per cent of total	5 4	19 8	19 1	16 8	17 6	11 3	10 1	100 0	
10. Per cent of average	37 6	138 6	133 9	117.7	122 9	78 8	70 4	700 0	100.0

B Percentages of average for each week

Week ended: (or row number)	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Total
1. March 27	44.2	135.7	132.8	131 2	137 8	69 0	49 4	700 1
2. April 3	63 9	154 8	116 9	105 9	115 9	70 6*	72 0	700 0
3. May 1	33.9*	151 4	129.4	116 6*	100 5	97 5	70 6	699 9
4. May 8	41.8	144 6*	161 0	129 4	105 6	61 4	56 2	700 0
5. May 15	28 8	135 1	131 8*	133 0	131 5	70 1	66 7*	700 0
6. May 22	25 3	150 1	141 4	107 0	125 4*	84 3	66 5	700 0
7. May 29	31 7	103 5	127 7	105 0	134 1	91 9	106 0	699 9
8 Arithmetic mean	38.5	139 3	134 4	118 3	122 0	77 8	69 6	699 9
9 Median	33.9	144 6	131 8	116 6	125 4	70 6	66 7	689 6
10. Index†	34.4	146 8	133 8	118.4	127.3	71 7	67 7	700.1

Source: Data of number of books from B. B. Downs, Librarian, The University of North Carolina
* Median
† Row 9 multiplied by correction factor 1 015081 = 700 ÷ 689 0

daily average as the percentage of average daily circulation during the week. Thus, in part A of the table, each number in row 8 is divided by 469.53 (= 3,286.71 ÷ 7), and the result set down as a percentage in row 10. Still other methods of stating the index are occasionally used. Sometimes the figures are put in terms of deviations above or below normal, either in absolute terms or in percentages.

Percentages of simple averages. It is a slight improvement of technique to express the circulation of each day of each week as a percentage of average circulation of that week. This gives each week equal importance in determining the index of variation, instead of extra weight being
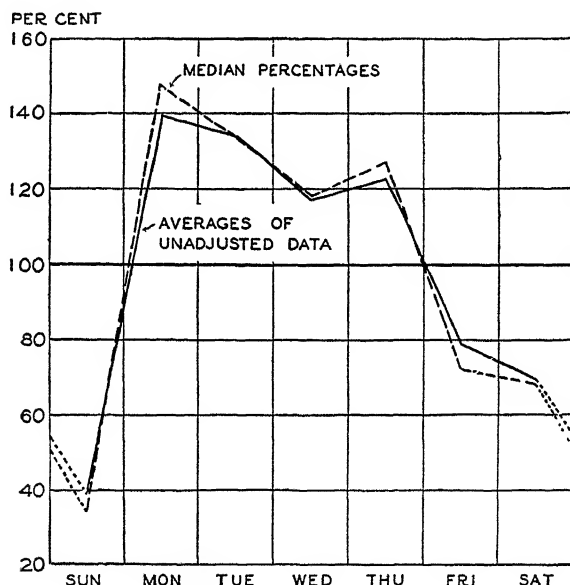
Chart 173. Two Measures of Typical Intra-Week Circulation of Books at the Reserve Room of the University of North Carolina, Spring Quarter, 1937. (Data of Table 104.)

accorded to weeks having large circulation. It might be thought offhand that such extra weight is highly desirable, but it does not necessarily follow that weeks of large circulation are weeks with typical circulation pattern. Furthermore, by putting the data in percentage form, we can more readily detect erratic variations from the typical weekly pattern. A study of such percentages for each day might lead one to select some average other than the arithmetic mean. Thus, in the present instance, use of the median gives the broken curve of Chart 173, which differs somewhat from the solid line representing the means of the unadjusted data. Computa-

tions are shown in part B of Table 104. Since medians were used, it was necessary to make a slight adjustment in order to have the index total 700 per cent and thus average 100 per cent.

Trend adjustment for averages. Whenever, as is usual, the data have a secular trend, any seasonal index following the simple methods just described will have an upward or downward bias, depending on the direction of the trend. Thus, if we were dealing with monthly data and if the trend were upward, each December would be higher than the preceding January by an amount equal to $\frac{11}{12}$ of the annual growth (assuming a linear trend), even if there were no genuine seasonal movement. Because of this fact the seasonal index, which is supposed to exhibit seasonal movements only, would slope upward; and if there were a true seasonal movement, the December index number would be too high relative to the January index number by $\frac{11}{12}$ of the annual growth.

In order to correct for trend, we can remove its influence either before or after averaging the data. The easiest method is to remove it after the data have been averaged. In Table 105, average electric power production per calendar day over the 10 years has been computed for each month, and is shown in row 11. Immediately below that are given the monthly trend values which refer to the 12-month period that is central in point of time of the entire series. (The trend equation is $Y_c = 192.86 + 1.48636X$, with origin at 1925–1926 and X units of one month.) These trend values may be thought of as the trend values of a typical year. Dividing the monthly averages by these trend values eliminates the effect of the secular trend, leaving the estimate of seasonal movements shown in row 13. The seasonal index will average approximately 100 per cent, since the total of row 12 is the same as that of row 11.

It is perhaps easier to make the trend correction in another fashion. Row 14 shows the cumulative trend increments as measured from the middle of a calendar year (that is, midway between June and July). Since the monthly trend increment (b in the trend equation) is 1.48636, the value referring to June is minus one-half of that amount, and July plus one-half that amount, so that July is larger than June by the amount of b; likewise August is larger than July by this amount, as is each month with respect to the preceding. The trend element present in row 14 is now subtracted and the results recorded in row 15. This procedure, however, does not give a seasonal index expressed as percentages. Were the seasonal estimates percentages of normal, they should average 100 per cent, and total 1200 per cent. Instead, they total 2314.2 (millions of kilowatt hours). Each of the numbers in row 15 can now be expressed as a percentage of average by dividing by 1.9285, since 192.85 is the average of the values in row 15. Mechanically it is easier to make the adjustment in a slightly

TABLE 105

COMPUTATION OF SEASONAL INDEX OF ELECTRIC POWER PRODUCTION, 1921–1930, BY ADJUSTMENT OF SIMPLE AVERAGES FOR SECULAR TREND

(Average per calendar day, millions of kilowatt hours)

Year (or description of row)	January	February	March	April	May	June	July	August	September	October	November	December	Total
1. 1921	114.2	113 1	109 5	108.0	105 3	108.1	105 5	110 0	112 5	115 3	121 3	123 2	
2. 1922	122 8	123 9	123 3	119 9	123 4	127 8	124 9	131.5	135 0	139 7	147 0	147 6	
3. 1923	152 8	154.0	151 9	148 6	149 5	150 0	145 5	149 6	150 1	158 5	160 4	159 5	
4. 1924	167.5	167.1	161.0	158 0	154.5	151 6	148.5	152 5	159 8	167 4	168 6	178 6	
5. 1925	179 8	178 6	173.9	172 7	169 0	174 9	173 8	176 3	183 2	191 9	192 9	198 5	
6. 1926	198.7	201.0	199 3	193.7	188.7	197 3	192 1	199 2	207 4	212 7	216 1	219 9	
7. 1927	220 3	220 2	220 6	216.1	212.9	216.4	208.9	215.9	220 2	223.6	229.2	232 6	
8. 1928	234 4	236 8	233 6	228 2	229 6	233 3	230.4	242 3	242 5	255 5	258 4	255 2	
9. 1929	265 8	265 4	257 8	262 7	260.8	258 9	260 4	269 5	268 7	280 9	274 7	274 6	
10. 1930	279 5	272 4	264 1	267.3	260 1	259.5	254 8	255 0	259 7	•264 4	256 4	261 5	
11. Average .. .	193 6	193 2	189 5	187 5	185.4	187 8	184 5	190 2	193 9	201 0	202 5	205 1	
12. Trend of average year .	184.7	186 2	187 7	189.1	190 6	192 1	193 6	195 1	196 6	198 1	199 5	201 0	1,200 3
13. Seasonal index*	104 8	103 8	101 0	99 2	97 3	97.8	95 3	97 5	98 6	101 5	101 5	102 0	
14. Trend increment .	-8 2	-6 7	-5 2	-3 7	-2 2	- 7	.7	2 2	3 7	5 2	6 7	8 2	
15. Row 11 minus row 14	201 8	199 9	194 7	191.2	187 6	188 5	183.8	188.0	190 2	195 8	195 8	196 9	2,314 2
16. Seasonal index†	104.6	103 7	101 0	99.1	97.3	97.7	95.3	97.5	98 6	101.5	101.5	102 1	1,199 9

* 100 × (row 11 ÷ row 12)
† Row 15 multiplied by 518538 = 1,200 0 ÷ 2,314 2

Source: Original data from United States Department of Commerce, *Survey of Current Business*, 1936 Supplement, p 85   For method of computing production per calendar day in 1930, see Table 78, first four columns.

different manner. Since the total of row 15 is $\dfrac{2,314.2}{1,200.0}$ of the value desired, it is necessary only to multiply each of the numbers in this row by $\dfrac{1,200.0}{2,314.2}$, or .518538. This is done in row 16. The method described in this paragraph ordinarily is not so logical as the method that eliminates the trend by dividing, since a time series is probably made up of $T \times C \times S \times I$ (Trend $\times$ Cycle $\times$ Seasonal $\times$ Irregular) rather than $T + C + S + I$. However, the discrepancy in the results usually, as in this case, is negligible.

The methods of adjustment for trend just described are not satisfactory when trend is an important component of the series. In years when trend is high, absolute values of the data will be large and these large values will have a disproportionate effect upon the monthly means. These methods are further limited in their usefulness to data the trend of which is linear.

Percentages of trend. A more satisfactory method of eliminating the disturbing element of trend is to compute percentages of monthly trend values before averaging the individual months.[1] Since one step in time series analysis frequently consists in computing the trend, the energy used in computing trend values is not wasted. To compute percentages of trend, we merely divide the original data by the trend values separately for each month and multiply by 100. The results of such computations for electric power production, 1921–1930, are shown in Table 106.

Since, however, this is a somewhat more refined method than the two previously described, it may be worth while to array the data before deciding upon the method of averaging. Inspection of the array of Table 107 in conjunction with Table 106 reveals that most of the high values occurred in 1923 or 1927, which were years of prosperity, while the low values tended to occur in the depression years of 1924 or 1930. This is not a surprising result; it indicates that the cyclical movements (though rather mild) are more powerful in this series than the irregular movements. The arithmetic mean is admirably adapted to averaging out random variations but not for taking care of cyclical movements, which, by their nature, are not distributed in random fashion. On the other hand, the median is better suited for averaging such data. Although other types of averages might also be satisfactory for this problem, the median is often selected because of its simplicity. The median of each column is recorded in row 11 of Table 107, and the index adjusted to total 1,200.0 in row 12. The

---

[1] This method is sometimes referred to as the *Falkner method.* See "The Measurement of Seasonal Variation," by Helen D. Falkner, *Journal of the American Statistical Association,* June 1924, pp. 167–179.

TABLE 106

Percentages of Trend of Electric Power Production, 1921–1930

(Original data: millions of kilowatt hours per calendar day)

Year	January	February	March	April	May	June	July	August	September	October	November	December
1921	109.4	106.8	102.0	99.2	99.1	96.6	93.1	95.8	96.7	97.9	101.7	102.0
1922	100.4	100.2	98.5	94.6	96.3	98.5	95.2	99.1	100.7	103.0	107.2	106.5
1923	109.1	108.8	106.1	102.8	102.1	101.7	97.7	99.4	98.8	103.3	103.5	102.0
1924	106.1	101.8	100.1	97.3	91.3	91.7	89.0	90.6	94.1	97.7	97.6	102.5
1925	102.3	100.7	97.3	95.8	95.0	95.5	94.1	94.7	97.6	101.5	101.2	103.3
1926	102.6	103.0	101.4	95.8	94.0	98.2	94.8	97.3	100.9	102.8	103.6	104.7
1927	104.2	103.4	102.9	100.1	97.9	98.9	94.8	97.3	98.6	99.5	103.3	102.1
1928	102.2	102.6	100.6	97.6	97.6	98.6	96.7	101.1	100.5	105.3	105.9	103.9
1929	107.6	106.8	103.1	104.4	103.0	101.7	101.7	104.7	103.7	107.8	104.8	104.2
1930	105.5	102.3	98.6	99.2	100.0	95.3	93.0	92.6	93.8	95.0	91.6	93.0

Source: Derived from data of Table 105

TABLE 107

Computation of Seasonal Index for Electric Power Production by Per Cent of Trend Method, 1921–1930

Rank (or description of row)	January	February	March	April	May	June	July	August	September	October	November	December	Total
1	109.4	108.8	106.1	104.4	103.0	101.7	101.7	104.7	103.7	107.8	107.2	106.5	
2	109.1	106.8	103.1	102.8	102.4	101.7	97.7	101.1	100.9	105.3	105.9	104.7	
3.	107.6	106.8	102.9	100.1	97.9	98.9	96.7	99.4	100.7	103.3	104.8	104.2	: :
4	106.1	104.8	102.0	99.2	97.6	98.6	95.2	99.1	100.5	103.0	103.6	103.9	.
5	105.5	103.4	101.4	99.2	96.3	98.5	94.9	97.6	98.8	102.8	103.5	103.3	::
6	104.2	103.0	100.6	97.8	96.0	98.2	94.8	97.3	98.6	101.5	101.7	102.5	
7.	102.6	102.6	100.1	97.6	95.4	96.6	94.1	95.8	97.6	99.5	101.3	102.1	. . .
8	102.3	102.3	98.6	97.3	94.6	95.5	93.1	94.7	97.9	97.9	101.2	102.0	
9	102.2	100.7	98.5	95.8	94.3	95.3	93.0	92.6	94.1	97.7	102.0	102.0	
10.	100.4	100.2	97.3	94.6	93.0	91.7	89.0	90.6	93.8	95.0	91.6	93.0	
11 Median	104.85	103.20	101.00	98.50	96.15	98.35	94.85	97.45	98.70	102.15	102.90	102.90	1,200.70
12 Seasonal index*	104.8	103.1	100.9	98.4	96.1	98.3	94.8	97.4	98.6	102.1	102.5	102.8	1,199.8

* Row 11 multiplied by correction factor. .999417 = 1,200 00 ÷ 1,200 70

Source: Table 106

adjustment was made by multiplying row 11 by 1,200.00 ÷ 1,200.70, or .999417.

This method of computing seasonals has something to commend it, since it is one of the simplest and easiest of methods. It is subject to the criticism, however, that it attempts to average out cyclical movements instead of eliminating them before the averaging process. An average cannot be expected to accomplish this if the cycles are pronounced, particularly if the period covered is short. Consequently it is most appropriately used when cyclical movements are unimportant relative to seasonal movements. Finally, the trend should be fitted to a period coinciding with that of the
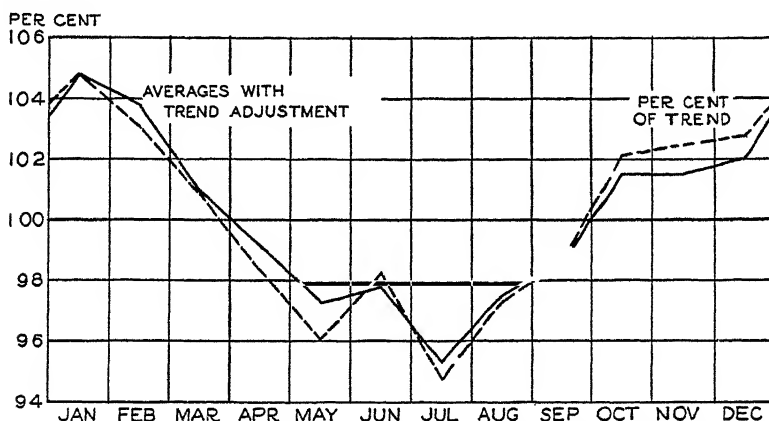
Chart 174. Seasonal Indexes of Electric Power Production per Calendar Day, by Two Methods, 1921–1936. (Data of Tables 105 and 107.)

seasonal measured. These limitations, of course, apply also to the methods previously discussed, which eliminate the trend after averaging the data. The use of this method, however, is not limited to data where the trend is linear.

In Chart 174 the seasonal index based on percentages of trend is compared with the index obtained by dividing by trend after averaging the raw data. Although the percentage of trend method gives more accurate results, the differences between the two indexes is rather slight. The greater accuracy of the per cent of trend method is probably due as much to the use of the median (instead of the arithmetic mean) as it is to the difference in the general method.

Percentages of 12-month moving average. In Chapter XV it was stated that moving averages iron out periodic movements if the moving average has the same number of months as the periodic movements sought to be eliminated. A 12-month moving average, therefore, should entirely eliminate seasonal movements if they are of constant pattern and ampli-

tude. Using the magazine advertising data, this moving average can be obtained by averaging together the first twelve months (January 1921 to December 1921, inclusive), then the second twelve months (February 1921 to January 1922, inclusive), and so on. Results are obtained most easily, however, as follows:

(1) Total separately the items for each calendar year and record these totals between June and July of their respective years. In Tables 108 and 109 these totals are starred

(2) From the 1921 calendar year figure subtract the figure for January 1921, add the figure for January 1922, and sub-total. From this sub-total subtract February 1921, add February 1922, and again sub-total. Continue this process. When the figure to be recorded between June and July 1922 is reached, it should check with the total already recorded as the total for the calendar year 1922. This procedure is carried through until all the moving total figures are obtained. The adding machine slip for the first 15 of the 12-month moving totals will appear as shown.

The first moving total figure was placed between June and July 1921, since the middle of that 12-month period is at midnight of June 30, 1921. Likewise, the second such period centers at the end of July 1921. But the monthly data are for calendar months, centering at the fifteenth of each

```
22,271.00*
 1,979.00-
 1,632.00
21,924.00 S
 1,981.00-
 1,768.00
21,711.00 S
 2,005.00-
 1,922.00
21,628.00 S
 2,099.00-
 2,171.00
21,700.00 S
 2,145.00-
 2,215.00
21,770.00 S
 1,933.00-
 2,046.00
21,883.00 S
 1,573.00-
 1,705.00
22,015.00 S
 1,402.00-
 1,566.00
22,179.00 S
 1,620.00-
 1,940.00
22,499.00 S
 1,824.00-
 2,470.00
23,145.00 S
 1,903.00-
 2,466.00
23,708.00 S
 1,807.00-
 2,464.00
24,365.00 S        Checked total,
 1,632.00-         starred in Table 108
 2,093.00
24,826.00 S
 1,768.00-
 2,301.00
25,359.00 S
```

Portion of Adding Machine Slip for Computation of Twelve-Month Moving Totals.

TABLE 108

Computation of Centered 12-Month Moving Average, and Percentages of Moving Average for United States Magazine Advertising, 1921–1922

Year and month (1)	Magazine advertising (thousands of lines) (2)	12-month moving total (3)	12-month moving average [Col 3 ÷ 12] (4)	2-month moving total of Col. 4 (5)	Centered 12-month moving average [Col 5 ÷ 2] (6)	Per cent of 12-month moving average [Col 2 ÷ Col 6] (7)
1921						
January	1,979					
February	1,981					
March	2,005					
April	2,099					
May	2,115					
June	1,933	22,271*	1,855 9			
July	1,573	21,924	1,827 0	3,682 9	1,841	85 4
August	1,402	21,711	1,809 2	3,636 2	1,818	77 1
September	1,620	21,628	1,802 3	3,611 5	1,806	89 7
October	1,824	21,700	1,808 3	3,610 6	1,805	101 1
November	1,903	21,770	1,814 2	3,622 5	1,811	105.1
December	1,807	21,883	1,823 6	3,637 8	1,819	99 3
1922						
January	1,632	22,015	1,834 6	3,658 2	1,829	89 2
February	1,768	22,178	1,848 2	3,682 8	1,841	96 0
March	1,922	22,499	1,874 9	3,723 1	1,862	103 2
April	2,171	23,144	1,928 7	3,803 6	1,902	114 1
May	2,215	23,708	1,975 7	3,904 4	1,952	113 5
June	2,046	24,365*	2,030 4	4,006 1	2,003	102 1
July	1,705	24,826	2,068 8	4,099 2	2,050	83 2
August	1,566	25,358	2,113 2	4,182 0	2,091	74 9
September	1,940	25,994	2,166 2	4,279 4	2,140	90 7
October	2,470	26,786	2,232 2	4,398 4	2,199	112 3
November	2,466	27,421	2,285 1	4,517 3	2,259	109 2
December	2,464	28,007	2,333 9	4,619 0	2,310	106 7

* Calendar year totals. See Table 109.

Source: See Table 109.

month. It is necessary, therefore, not only to convert the moving totals into moving averages, but to center them at the middle of each month. Two procedures will be explained for accomplishing this result.

The longer, although more easily understood, method is illustrated for 1921 and 1922 in Table 108. Divide each of the moving totals by 12; or, to save time, multiply by .0833333, the reciprocal of 12. This process obtains the moving averages. They are centered by taking a 2-month moving average of the 12-month moving averages. This is done by taking a 2-month moving total of column 4 and dividing by 2. The results, recorded in column 6, may be called a centered 12-month moving average.[2]

A short cut method which gives precisely the same results is illustrated in Table 109. In columns 4 and 5 of this table we center the moving totals, rather than the moving averages, and convert them into moving averages by multiplying by .0416667, the reciprocal of 24. The merit of this method is that one column of divisions (multiplications by reciprocals) is eliminated.

These results, together with the original data, are plotted in Chart 175A

The next step is to divide the original data by the centered 12-month moving average in order to express them as percentages of this average. This is done in Table 109, with results recorded in column 6. The values are the same as those of Table 108, column 7. These results are shown by the solid line of Chart 175B. The logic of the procedure is as follows: Time series are assumed to be composed of $T \times C \times S \times I$ (Trend $\times$ Cycle $\times$ Seasonal $\times$ Irregular). The 12-month moving average is a rough estimate of $T \times C$ because the 12-month average smooths out seasonal movements and, for the most part, irregular movements, since the latter are largely movements of small amplitude and short duration.[3] If now we divide the original data by the 12-month moving average, we have an estimate of the seasonal and irregular movements combined:

$$\frac{T \times C \times S \times I}{T \times C} = S \times I.$$

It is well at this point to take note of the progress that has been made by comparing sections A and B of Chart 176. Section A is a rearrange-

---

[2] Some statisticians do not consider the added accuracy obtained by centering to be worth the added effort. The difference in results in the seasonal index is usually very slight.

[3] Actually the irregularities are not entirely smoothed out; on the other hand, some of the cyclical movements are partly smoothed out, so that the amplitude of the cyclical movements is slightly reduced. Consequently some statisticians smooth the moving average curve further by inspection and also alter it slightly, particularly at cyclical peaks and troughs, so as to obtain a better estimate of $T \times C$. These adjusted data are then used in subsequent steps for obtaining the seasonal index

TABLE 109

SHORT METHOD OF COMPUTING CENTERED 12-MONTH MOVING AVERAGE, AND OF
PERCENTAGES OF MOVING AVERAGE FOR UNITED STATES MAGAZINE
ADVERTISING, 1921–1930

Year and month (1)	Magazine advertising (thousands of lines) (2)	12-month moving total (3)	2-month moving total of Col. 3 (4)	Centered 12-month moving average [Col 4 ÷ 24] (5)	Per cent of 12-month moving average [Col 2 ÷ Col. 5] (6)
1921					
January	1,979				.
February	1,981				.
March	2,005	.	.		.
April	2,099		.	.	.
May	2,145	.	..	.	.
June	1,933	22,271*	...	.	.
July	1,573	21,924	44,195	1,841	85 4
August	1,402	21,711	43,635	1,818	77 1
September	1,620	21,628	43,339	1,806	89 7
October	1,824	21,700	43,328	1,805	101 1
November	1,903	21,770	43,470	1,811	105 1
December	1,807	21,883	43,653	1,819	99 3
1922					
January	1,632	22,015	43,898	1,829	89 2
February	1,768	22,179	44,194	1,841	96 0
March	1,922	22,499	44,678	1,862	103 2
April	2,171	23,145	45,644	1,902	114 1
May	2,215	23,708	46,853	1,952	113.5
June	2,046	24,365*	48,073	2,003	102 1
July	1,705	24,826	49,191	2,050	83 2
August	1,566	25,359	50,185	2,091	74 9
September	1,940	25,994	51,353	2,140	90 7
October	2,470	26,786	52,780	2,199	112 3
November	2,466	27,421	54,207	2,259	109 2
December	2,464	28,007	55,428	2,310	106 7
1923					
January	2,093	28,452	56,459	2,352	89.0
February	2,301	28,750	57,202	2,383	96 6
March	2,557	29,087	57,837	2,410	106 1
April	2,963	29,490	58,577	2,441	121.4
May	2,850	29,924	59,414	2,476	115.1
June	2,632	30,233*	60,157	2,507	105 0
July	2,150	30,421	60,654	2,527	85.1
August	1,864	30,677	61,098	2,546	73 2
September	2,277	31,035	61,712	2,571	88 6
October	2,873	31,345	62,380	2,599	110 5
November	2,900	31,483	62,828	2,618	110.8
December	2,773	31,681	63,164	2,632	105.4

TABLE 109 (Continued)

SHORT METHOD OF COMPUTING CENTERED 12-MONTH MOVING AVERAGE, AND OF
PERCENTAGES OF MOVING AVERAGE FOR UNITED STATES MAGAZINE
ADVERTISING, 1921–1930

Year and month	Magazine advertising (thousands of lines)	12-month moving total	2-month moving total of Col. 3	Centered 12-month moving average [Col 4 − 24]	Per cent of 12-month moving average [Col 2 ÷ Col 5]
(1)	(2)	(3)	(4)	(5)	(6)
1924					
January	2,281	31,548	63,229	2,635	86 6
February	2,557	31,503	63,051	2,627	97 3
March	2,915	31,468	62,971	2,624	111 1
April .	3,273	31,374	62,842	2,618	125 0
May .	2,988	31,374	62,748	2,615	114 3
June .	2,830	31,442*	62,816	2,617	108 1
July .	2,017	31,272	62,714	2,613	77.2
August	1,819	31,228	62,500	2,604	69 9
September	2,242	31,025	62,253	2,594	86 4
October .	2,779	30,703	61,728	2,572	108 0
November	2,900	30,569	61,272	2,553	113.6
December.	2,841	30,374	60,943	2,539	111.9
1925					
January .	2,111	30,425	60,799	2,533	83.3
February .	2,513	30,484	60,909	2,538	99 0
March. .	2,712	30,727	61,211	2,550	106 4
April . .	2,951	31,010	61,737	2,572	114.7
May. .	2,854	31,354	62,364	2,599	109 8
June	2,635	31,473*	62,827	2,618	100 6
July	2,068	31,747	63,220	2,634	78 5
August . .	1,878	32,088	63,835	2,660	70.6
September.	2,485	32,406	64,494	2,687	92 5
October .	3,062	32,798	65,204	2,717	112.7
November ..	3,244	33,180	65,978	2,749	118.0
December...	2,960	33,569	66,749	2,781	106 4
1926					
January.	2,385	33,869	67,438	2,810	84 9
February . . .	2,854	34,172	68,041	2,835	100 7
March......	3,030	34,518	68.690	2,862	105.9
April.	3,343	34,900	69,418	2,892	115 6
May	3,236	35,242	70,142	2,923	110 7
June	3,024	35,491*	70,733	2,947	102 6
July..	2,368	35,642	71,133	2,964	79 9
August........	2,181	35,793	71,435	2,976	73 3
September..	2,831	36,018	71,811	2,992	94.6
October	3,444	36,172	72,190	3,008	114.5
November....	3,586	36,513	72,685	3,029	118.4
December... .	3,209	36,501	73,014	3,042	105.5

TABLE 109 (Continued)

SHORT METHOD OF COMPUTING CENTERED 12-MONTH MOVING AVERAGE, AND OF
PERCENTAGES OF MOVING AVERAGE FOR UNITED STATES MAGAZINE
ADVERTISING, 1921–1930

Year and month (1)	Magazine advertising (thousands of lines) (2)	12-month moving total (3)	2-month moving total of Col 3 (4)	Centered 12-month moving average [Col 4 ÷ 24] (5)	Per cent of 12-month moving average [Col 2 ÷ Col. 5] (6)
1927					
January	2,536		73,054	3,044	83 3
February	3,005	36,553	73,154	3,048	98 6
March	3,255	36,601	73,133	3,047	106 8
April	3,497	36,532	73,031	3,043	114 9
May	3,577	36,499	72,985	3,041	117.6
June	3,012	36,486	72,939	3,039	99 1
July	2,420	36,453*	72,817	3,034	79 8
August	2,229	36,364	72,569	3,024	73.7
September	2,762	36,205	72,367	3,015	91.6
October	3,411	36,162	72,502	3,021	112 9
November	3,573	36,340	72,538	3,022	118 2
December	3,176	36,198	72,445	3,019	105 2
		36,247			
1928					
January	2,447		72,657	3,027	80 8
February	2,846	36,410	72,749	3,031	93 9
March	3,212	36,339	72,721	3,030	106 0
April	3,675	36,382	72,852	3,036	121 0
May	3,435	36,470	72,853	3,036	113 1
June	3,061	36,383	72,762	3,032	101.0
July	2,583	36,379*	72,995	3,041	84.9
August	2,158	36,616	73,544	3,064	70 4
September	2,805	36,928	74,245	3,094	90 7
October	3,499	37,317	75,041	3,127	111.9
November	3,486	37,724	75,888	3,162	110 2
December	3,172	38,164	76,814	3,201	99.1
		38,650			
1929					
January	2,684		77,581	3,233	83.0
February	3,158	38,931	78,134	3,256	97 0
March	3,601	39,203	78,763	3,282	109.7
April	4,082	39,560	79,381	3,308	123.4
May	3,875	39,821	79,984	3,333	116.3
June	3,547	40,163	80,769	3,365	105.4
July	2,864	40,606*	81,033	3,376	84.8
August	2,430	40,427	80,720	3,363	72 3
September	3,162	40,293	80,401	3,350	94 4
October	3,760	40,108	80,011	3,334	112 8
November	3,828	39,903	79,570	3,315	115.5
December	3,615	39,667	79,141	3,298	109 6
		39,474			

TABLE 109 (Continued)

SHORT METHOD OF COMPUTING CENTERED 12-MONTH MOVING AVERAGE, AND OF PER-
CENTAGES OF MOVING AVERAGE FOR UNITED STATES MAGAZINE
ADVERTISING, 1921–1930

Year and month (1)	Magazine advertising (thousands of lines) (2)	12-month moving total (3)	2-month moving total of Col. 3 (4)	Centered 12-month moving average [Col 4 ÷ 24] (5)	Per cent of 12-month moving average [Col. 2 ÷ Col 5] (6)
1930					
January. .	2,505	39,061	78,535	3,272	76.6
February . .	3,024	38,688	77,749	3,240	93.3
March	3,416	38,124	76,812	3,201	106 7
April . .	3,877	37,385	75,509	3,146	123 2
May.. . ..	3,639	36,599	73,984	3,083	118 0
June .	3,354	35,804*	72,403	3,017	111 2
July	2,451	. .	...	. .	...
August .	2,057	. .	...	. .	...
September	2,598	. .	...	...	...
October . .	3,021	...	...	. .	...
November .	3,042	...	...	...	...
December .	2,820	.	.	.	...

* Calendar year totals
Source United States Department of Commerce, *Survey of Current Business*, October 1933, p. 20;
1936 Supplement, p 24, December 1936, p 25; May 1937, p. 26

ment of the data of the solid line of Chart 175A, while section B is a re-arrangement of the data of the solid line of Chart 175B. First, it is to be noted that the seasonal movements appear to be more regular in section B. This is because section A contains trend and cyclical movements, while these have been largely eliminated in section B. Secondly, a careful inspection of section B indicates that the seasonal pattern has not been uniform throughout the entire period. The amplitude of the seasonal swing is larger in the more recent years, the April peak having become relatively more pronounced. Although this change did not take place abruptly, nevertheless, the pattern seems to be fundamentally different after 1929. Consequently it seems advisable to compute two separate indexes, one based on 1922–1929 percentages, and the other on 1930–1935 percentages.[4]

The procedure is now similar to that followed in obtaining a seasonal index from percentages of trend. In that case the problem was to average

[4] A method for obtaining a progressively changing or "moving" seasonal is explained in Chapter XVIII.

out cyclical and irregular movements.   In the present instance it is mainly
irregular movements which must be removed by averaging.

In order to study the distribution more carefully, section A of Chart
177 is made.   Each circle or dot represents one of the percentages of
Table 110, which is a rearrangement of data in the last column of Table
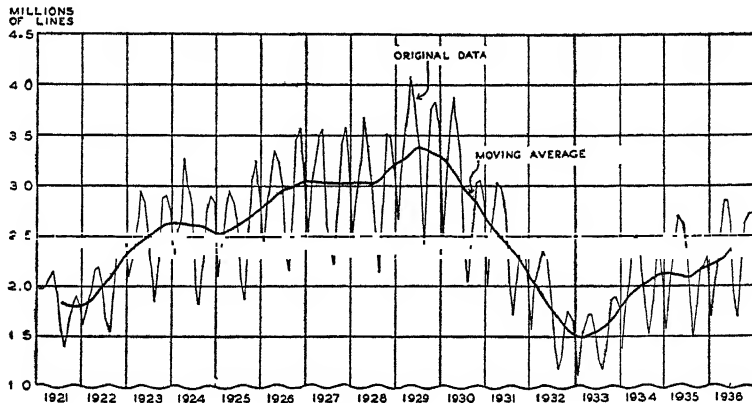
Chart 175A.   United States Magazine Advertising and 12-Month Moving Average,
1921-1936.   (1921-1929 data are from Table 109; 1930-1936 data are based on figures
of Table 116.)
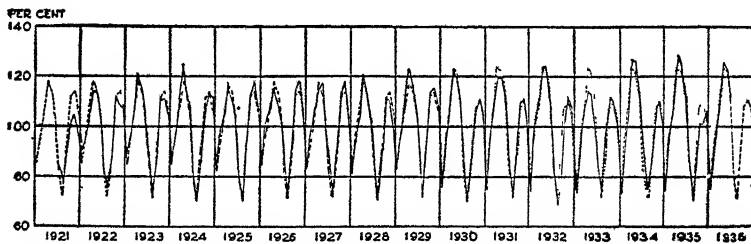
Chart 157B.   Percentages of 12-Month Moving Average of United States Magazine
Advertising and Seasonal Indexes, 1921-1936.   (Percentages of moving average are
represented by solid line, seasonal index by dotted line.   Based on figures of Tables 109
and 116.)

109.   The scatter is in the main due to irregular movements, but it is also
partly due to lack of complete homogeneity.   Use of a 12-month moving
average cannot completely eliminate cycles; it will not reach up into the tip
of their peaks or drop into the bottom of the troughs.   A tabulation of the
number of times that each year includes one of the two highest percentages
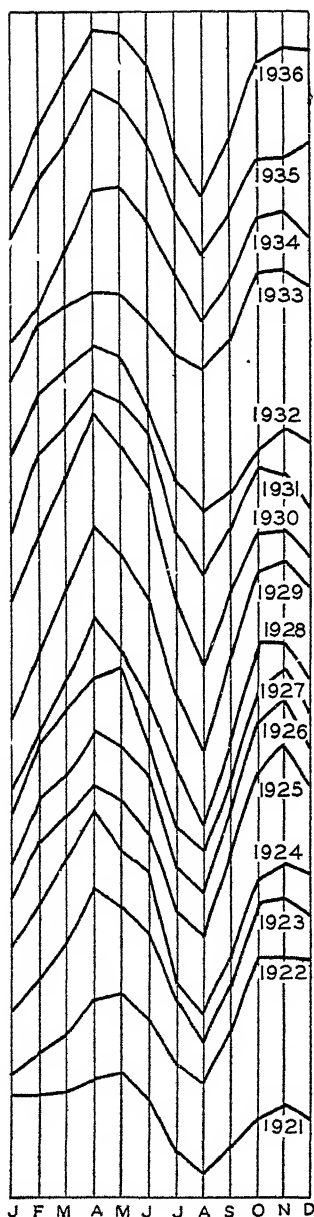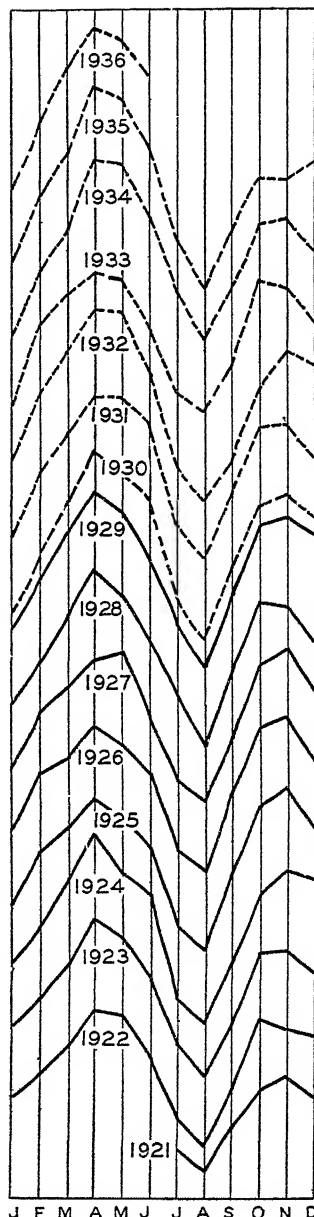
Chart 176. United States Magazine Advertising: A. Unadjusted, and B. Percentages of 12-Month Moving Average, 1921-1936. (For purposes of comparison the different curves have been plotted close together on the same chart instead of on separate charts. Each curve is plotted to the same vertical scale, but at a different level. This arrangement is sometimes referred to as a multiple axis chart For source of data, see Chart 175.)

for any monthly column of Table 110, or one of the two lowest, gives these results:

Year	Two highest	Two lowest
1922	2	4
1923	2	2
1924	4	4
1925	1	4
1926	4	2
1927	4	2
1928	1	5
1929	6	1
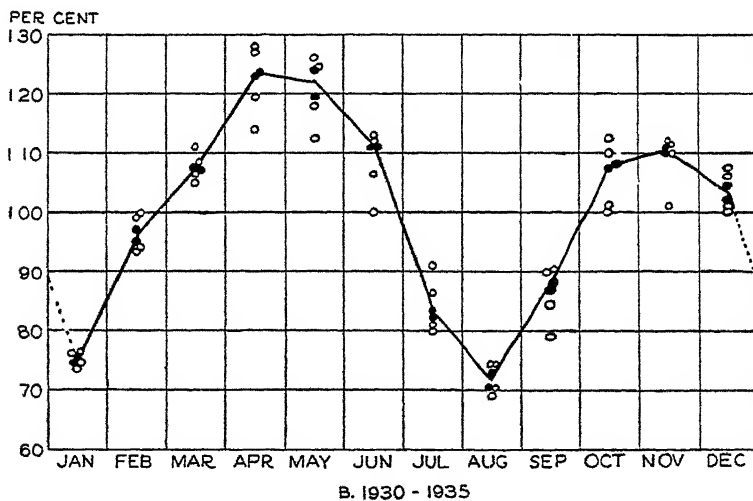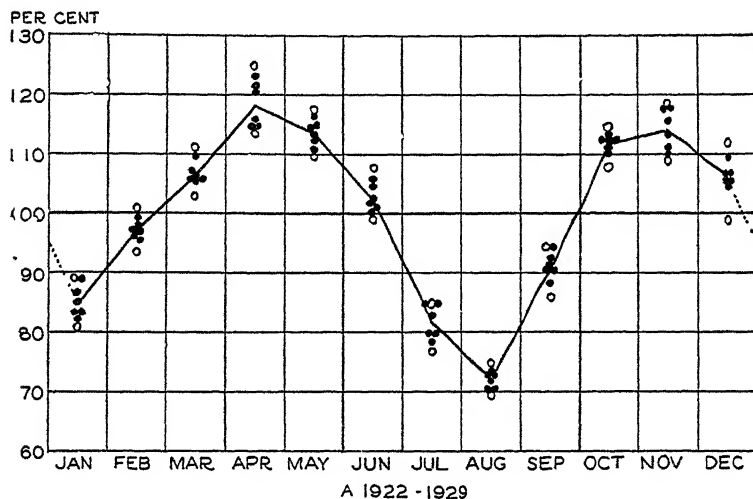
Chart 177.   Arrayed Percentages of 12-Month Moving Average and Seasonal Index, United States Magazine Advertising: A. 1922-1929, B. 1930-1935.   (1922-1929 data are from Table 111, 1930-1935 data are based on figures of Table 116.)

TABLE 110

Percentages of 12-month Moving Averages of United States Magazine Advertising, 1921–1929

Year	January	February	March	April	May	June	July	August	September	October	November	December
1921					113.5	102.1	85.4	77.1	89.7	101.1	105.1	99.3
1922	80.2	96.0	103.2	114.1	115.1	105.0	83.2	74.9	90.7	112.3	109.2	106.7
1923	89.0	96.6	106.1	121.4	114.3	108.1	85.1	73.2	88.6	110.5	110.8	105.4
1924	86.6	97.3	111.1	125.0	109.8	106.1	77.2	69.0	86.4	108.1	113.6	111.9
1925	84.9	99.0	106.4	114.7	110.7	102.6	78.5	70.6	92.5	112.7	118.0	104.4
1926	83.3	100.7	105.9	115.6	117.6	99.1	79.8	73.7	94.6	114.5	118.4	105.5
1927	83.3	98.6	106.0	114.9	113.1	101.0	79.8	73.7	91.6	112.9	118.2	105.2
1928	80.8	93.9	100.8	121.0	117.6	101.0	84.9	70.4	90.7	111.9	110.2	99.1
1929	83.0	97.0	109.7	123.4	116.3	105.4	81.8	72.3	94.4	112.8	115.5	109.6

Source: Table 109

TABLE 111

Arrays of Percentages of Centered 12-month Moving Averages and Computation of Seasonal Index for United States Magazine Advertising, 1922–1929

Rank (or description of row)	January	February	March	April	May	June	July	August	September	October	November	December	Total
1	89.2	100.7	111.1	125.0	117.6	108.1	85.1	74.9	94.6	114.5	118.4	111.9	· ·
2	89.0	99.0	109.7	123.4	116.3	105.4	84.9	73.7	94.4	112.9	118.2	109.6	
3	86.6	98.6	106.8	121.4	115.1	105.0	84.8	73.3	92.5	112.8	118.0	106.7	
4	84.9	97.3	106.4	121.0	114.3	102.6	83.2	73.2	91.6	112.7	115.5	106.4	
5	83.3	97.0	106.1	115.6	113.5	102.1	79.9	72.3	90.7	112.3	113.6	105.5	
6	83.3	96.6	106.0	114.9	113.1	101.0	79.8	70.6	90.7	111.9	110.8	105.4	
7	83.0	96.0	105.9	114.7	110.7	101.0	78.5	70.4	88.6	110.5	110.2	105.2	
8	80.8	93.9	103.2	114.1	109.8	99.1	77.2	69.9	86.4	108.0	109.2	105.2	
9 Modified mean	85.0	97.4	106.8	118.5	113.8	102.8	81.8	72.2	91.4	112.2	114.4	106.5	1,202.8
10 Seasonal index*	84.8	97.2	106.6	118.2	113.5	102.6	81.6	72.0	91.2	111.9	114.1	106.3	1,200.0

* Row 9 multiplied by correction factor .997672 = 1,200.0 ÷ 1,202.8
Source. Table 110

482

It is significant that five low values occurred in 1928, a year of depression for magazine advertising, and six high values occurred in 1929, a year of prosperity. The above tabulation does not tell the full story. For instance, four highs and four lows occurred in 1924, but advertising was high in the spring of 1924 and low in the fall. On account of the limited number of observations, an extreme deviation, whether an irregular movement or one of the systematic variation just described, exercises an undue influence on an arithmetic mean. Probably, therefore, it is well to eliminate the more extreme deviations before computing averages of the different months.

There are two ways of deciding what items to eliminate. One way is to consider each array of section A of Chart 177 separately and to eliminate items that appear to be unusually high or low, perhaps studying each large deviation individually and eliminating those for which a special circumstance can be discovered. If this method is followed, one array might use an average of all items, another might employ the median, a third the central six items, a fourth all items except the two highest, etc. These averages might be called *modified means*. On account of the extreme subjectivity of the method, it is dangerous unless the statistician possesses a high order of knowledge and judgment.

A more objective method, and the one which is recommended here, employs uniform modified means. Each month is treated exactly alike, and an average is computed for the central items in an array. The number of items to exclude is determined by inspection of a diagram like section A of Chart 177. Although this decision is subjective, a curb is put on the statistician's exercise of judgment by the necessity of treating each array alike. No generally accepted rule concerning the number of items to exclude can be laid down, but very often it will be found advisable to exclude roughly the highest and lowest ten per cent of each array. The number to exclude bears a relationship to the duration of the cyclical movements of the series: the shorter their duration, the larger the proportion which should be excluded, other factors being equal.

Another way of explaining the desirability of this type of average is as follows: A mean becomes more reliable as the sample becomes larger, provided the data are homogeneous. As the proportion of items averaged from a monthly array is increased, the mean therefore becomes progressively more reliable up to a certain point, when the increasing heterogeneity of these data reverses the tendency. The trick, then, is to include as many items as possible without including too many that are too unrepresentative. In the present instance the middle six out of eight were averaged. These are shown as dots on section A of Chart 177 and the excluded extremes left as hollow circles.

Table 111 is a computation table similar to Table 107. In it the data of Table 110 for calendar years 1922–1929 have been arrayed by months. The items to be excluded from our averages have been set off by horizontal lines. The modified means are shown in row 9. These total 1,202.8, and are adjusted to total 1,200.0 by multiplying by the correction factor .997672, which is 1,200.0 ÷ 1,202.8.[5]

In similar fashion an index based upon medians of the 1930–1935 percentages has been computed, and is shown graphically by section B of Chart 177. It should be compared with section A of the same chart. Not only is the seasonal pattern different, but the data conform less consistently to the seasonal pattern, as can be seen from the dispersion of the circles and dots. Irregularity of economic events has been a characteristic of the years following 1929.

A graphic approach to seasonal measurement. The use of the 12-month moving average involves a considerable amount of labor. A graphic method is available, however, which is similar in logic to that method, and which produces reasonably accurate results if skillfully handled.[6]

First the data are plotted on semi-logarithmic paper. It is advisable to select accurately ruled paper with a large vertical scale (about 8 inches to a cycle), since measurements are to be made from the chart. Needless to say, the plotting must be accurate. Next is plotted, by inspection, an estimate of the combined trend and cyclical movements. If desired, annual averages may first be plotted in the middle of each year as a partial guide. The plotting of the Trend × Cycle estimate is largely subjective, however, and requires a high order of judgment on the part of the statistician. Chart 178 illustrates the first two steps of this method. The similarity of the broken line of this chart to the 12-month moving average of Chart 175A should be noted, as well as the differences between them. The differences are mainly: (1) the freehand curve is smoother; (2) it is more flexible, dipping further into the troughs and reaching higher into the peaks. It is largely on account of the apparently more faithful representation of Trend × Cycle by this line that the graphic method is claimed by some authorities to be superior to the 12-month moving average method.

The next step is to compute graphically the ratio of the original data to the estimate of Trend × Cycle. First, on a small card or piece of paper,

---

[5] It is possible to introduce a short cut into the computation of the seasonal index by using percentages of 12-month moving totals rather than 12-month moving averages. If this is done, the total of the modified means will be in the neighborhood of 100.00, and the correction factor will be in the neighborhood of 12 rather than 1.

[6] See "A Graphic Method of Measuring Seasonal Variation," by William A. Spurr *Journal of the American Statistical Association*, Vol. 32, June 1937, pp. 281–289.

draw a broken line perpendicular to its edge and label this line Trend ×
Cycle. Next, place the edge of this card on the chart at January 1921,
with the Trend × Cycle mark directly on the broken line of the chart.
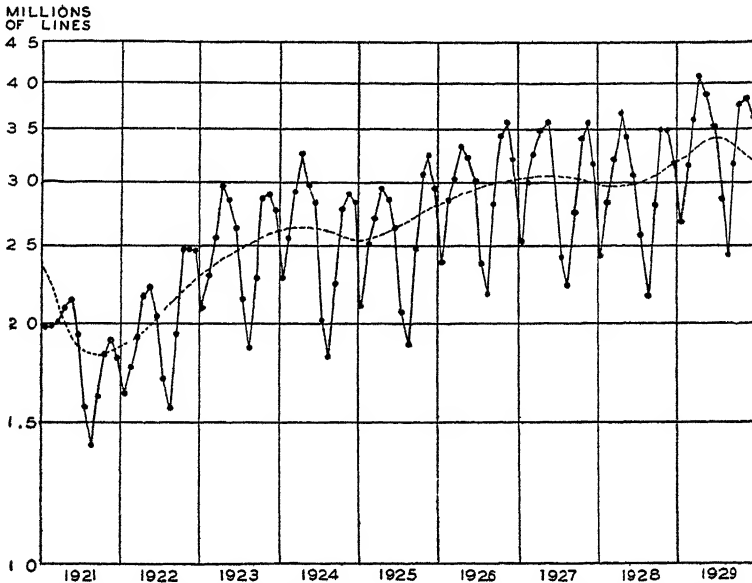
Chart 178. United States Magazine Advertising and Estimate of Trend × Cycle
Movements, 1921-1929; Logarithmic Vertical Scale. (Data of Table 109, Col. 2)

Now, put a small mark on the edge of the card at the point which coincides
with the plotted point on the solid line for January 1921. Repeat this
process for January of each year. We now have ten
lines running to the edge of the card, a fairly long
one representing the Trend × Cycle and nine
shorter ones representing the different Januaries.
Since January advertising is always rather low,
each of these small marks lies below the Trend
× Cycle base line.

Chart 179. Graphic
Observations of January
Seasonal Variation of
United States Magazine
Advertising for Years
1921-1929. (Readings
are from Chart 178.)
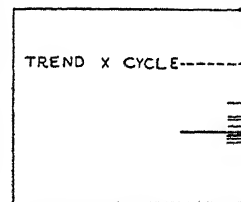
The January measurements are indicated on Chart
179. The distances between the broken line and the
various short lines are graphic representations of
estimates of the January seasonal ratio. The dif-
ferent estimates seem to group around some central
point; consequently an average estimate can be
marked on the card with reasonable accuracy. This average, which is a
modified mean, is the long solid line of the chart. In order to obtain

a numerical value for this average, we must place the card on a logarithmic scale of the same size as that of Chart 178 and with the dotted line on the scale value representing 100, and read the scale value corresponding to the average line on the card. In this instance the reading is 84.1.

A similar set of measurements is now made for each month, with results shown graphically in Chart 180. The sum of the readings will not total exactly 1,200, and so an adjustment must be made, as in Table 112.

TABLE 112

Computation of Seasonal Index by Graphic Method for United States Magazine Advertising, 1921–1929

Month	Chart reading	Seasonal index*
January . . .	84.1	84 0
February	98 0	97 9
March	107 0	106 9
April . ..	117 0	116 9
May	112 5	112 4
June	102 5	102.4
July	82 0	81.9
August	72 8	72.7
September . .	90.5	90 4
October . ..	112 8	112 7
November.. ..	115.8	115.7
December .	106 5	106.4
Total . . .	1,201.5	1,200.3

* Correction factor 99875 = 1,200 0 ÷ 1,201 5.

Link relative method. Though not so simple as the moving average method or so easily adapted to complex types of seasonal movement, the actual computations of the link relative method are much less extensive. At one time it was probably the most widely used method. This method is based upon the averaging of link relatives. A *link relative* is the value for one month expressed as a percentage of the preceding. Thus the following values are link relatives:

$$\frac{\text{Jan}}{\text{Dec}}, \frac{\text{Feb}}{\text{Jan}}, \frac{\text{Mar}}{\text{Feb}}, \text{etc.}$$

Although subject to variation in detail, the steps involved in the calculations may be summarized as follows:

1. *Express each month as a percentage of the preceding month.* This is done in Table 113.
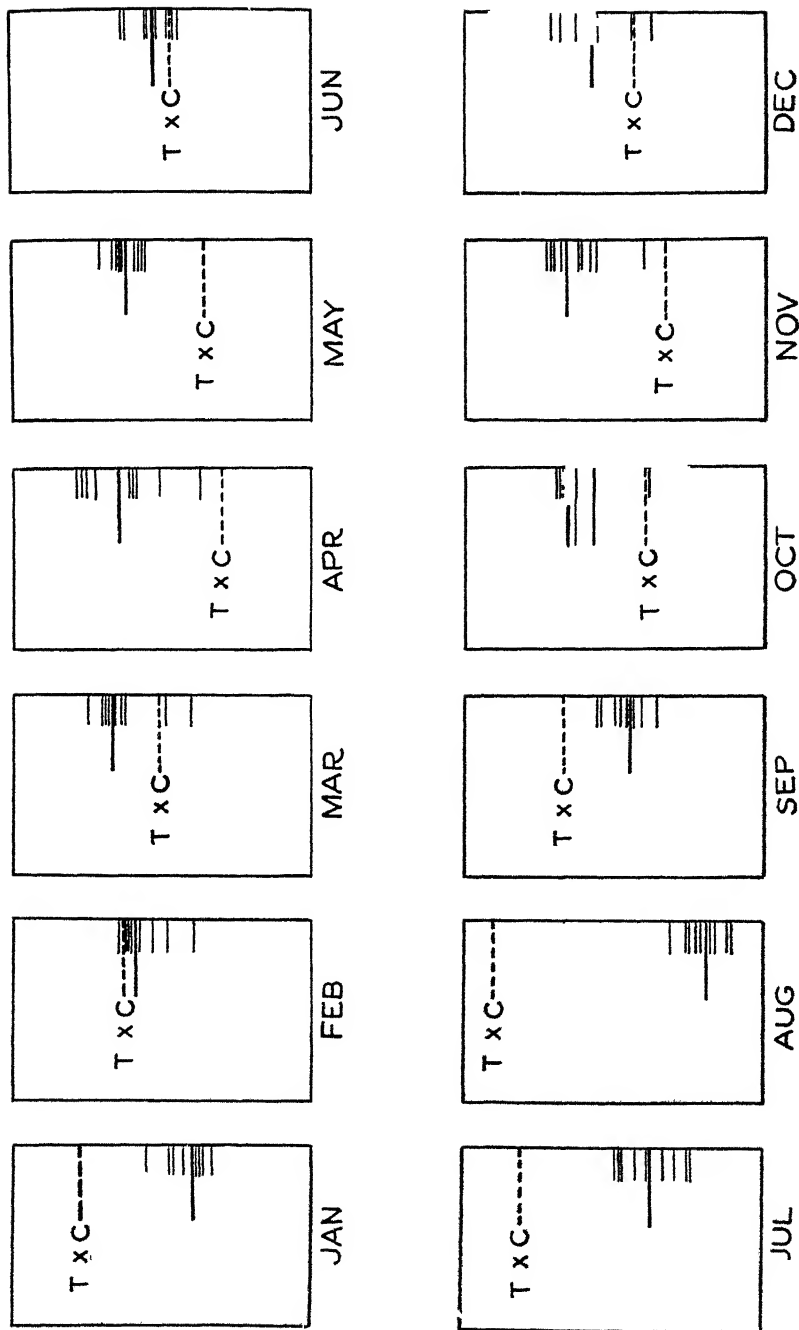
Chart 180. Graphic Observations of Seasonal Movements of United States Magazine Advertising, 1921-1929. (Readings are from Chart 178.)

487

The labor of these computations can be lightened if the following procedure is adopted. First, place the last number of the entire series in the calculating machine and divide by the next to the last. In the present instance

$$\frac{\text{December 1929}}{\text{November 1929}} = \frac{3,615}{3,828} = 94.4.$$

When this result is obtained, clear the dials, but not the keyboard. Thus 3,828 is still on the keyboard and may now be put in the machine preparatory to being divided by 3,760, the October 1929 value. By this procedure a number need be placed on the keyboard only once, whereas if the work
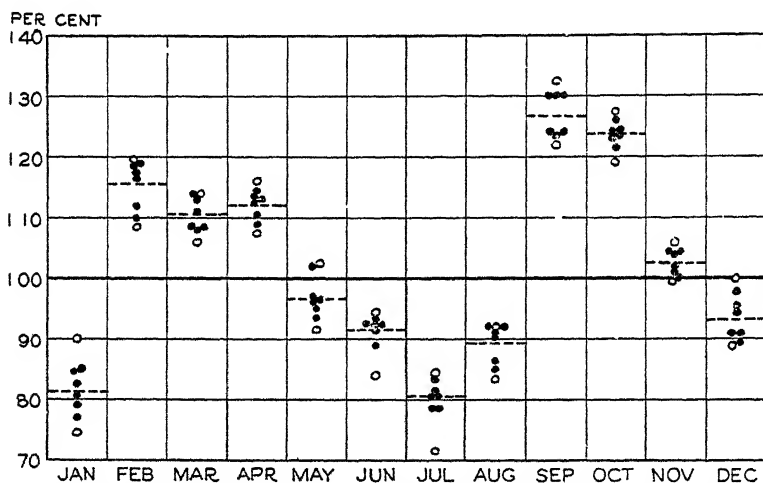
Chart 181. Arrayed Link Relatives and Modified Means, United States Magazine Advertising, 1922-1929. (Data of Table 114)

proceeded from earliest month to latest, each number would have to be set up twice.

2. *Average the link relatives for each month separately.* The average employed is usually the median but may be a modified mean. In the present instance the mean of the middle six of the arrays of eight items was selected. See Chart 181 and Table 114.

3. *Chain the link relatives to the preceding month by successive multiplication.* This is done in row 10 of Table 114. January is arbitrarily taken as 100 per cent, and the other numbers are thus percentages of the January value. The last entry in this row, 111.89, is not the total of the row, but is the new January chain relative obtained by multiplying 137.46 by .814. (When multiplying percentages, remember that 115.6 per cent × 110.5 per cent = 1.156 × 1.105 = 1.2774, or 127.74 per cent.)

TABLE 113

UNITED STATES MAGAZINE ADVERTISING LINK RELATIVES, 1922–1929

Year	January	February	March	April	May	June	July	August	September	October	November	December
1922	90 3	108 3	108 7	113 0	102 0	92 4	83 3	91 8	123 9	127 3	99 8	99 9
1923	84 9	109.9	111 1	115 9	96 2	92 4	81 7	86 7	122 2	126 2	100 9	95 6
1924	82 3	112 1	111 0	112 3	91 3	94 7	71 5	90 2	123 3	124 0	104 4	98 0
1925	74 3	119 7	107 9	108 8	96 7	92 3	78 5	90 8	132 3	123 2	105 9	91 2
1926	80 6	119 7	106 2	110 3	96 2	93 4	78 3	92 1	129 8	121 7	104 1	89 5
1927	79 0	118.5	108 3	107 4	102 3	84 2	80 3	92 1	123 9	123 5	104 7	88 9
1928	77 0	116 3	109 0	114 4	93 5	89 1	84 4	83 5	130 0	124 7	99 6	91 0
1929	84 6	117 7	111 0	113 4	94 9	91 5	80 7	84 8	130 1	118 9	101 8	94 4

Source: Table 109

TABLE 114

COMPUTATION OF LINK RELATIVE SEASONAL INDEX FOR UNITED STATES MAGAZINE ADVERTISING, 1922–1929

Rank (or description of row)	January	February	March	April	May	June	Jul.	August	September	October	November	December	Total
1.	90 3	119 7	114 0	115 9	102 3	94 7	81 4	92 1	132 3	127 3	105 9	99 9	
2.	84 9	119 0	114 0	114 4	102 0	93 4	83 3	92 1	130 1	126 2	104 7	98 0	
3.	84 6	118 5	112 9	113 4	96 8	92 4	81 7	91 8	130 0	124 7	104 4	95 4	
4.	82 3	117 7	111 1	113 0	96 7	92 4	80 7	90 8	129 8	124 0	104 1	94 4	
5.	80 6	116 3	108 7	110 3	96 2	92 3	80 3	90 2	123 9	123 5	100 9	91 2	
6.	79 0	112 1	108 3	108 8	94 9	91 5	78 5	85 7	123 3	123 2	100 9	91 0	
7.	77 0	109 9	107 9	108 8	93 5	89 1	78 3	84 8	123 3	121 7	99 8	89 5	
8.	74 3	108 3	106 2	107 4	91 3	84 2	71 3	83 5	122 2	118 9	99 6	88 9	
9. Average of middle 6 items	81 4	115 60	110 5	112 0	96 7	91 8	80 5	89 4	126 8	123 9	102 6	93 3	
10. Uncorrected chain relative	100 00	115 60	127 74	143 07	138 35	127 01	102 24	91 40	115 90	143 60	147 33	137 46	*111 89*
11. Less, trend correction†		99	1 98	2 97	3 96	4 95	5 94	6 94	7 93	8 92	9 91	10 90	*11 89*
12. Chain relatives corrected for trend‡	100 00	114 61	125 76	140 10	134 39	122 06	96 30	84 46	107 97	134 68	137 42	126 56	1,424 31
13. Seasonal index‡	84 3	96 6	106 0	118 0	113 2	102 8	81 1	71 2	91 0	113 5	115 8	106 6	1,200 1

* 111 89 = 137 46 × 8 14
† Successive increments of 11 89 ÷ 12 = .9908
‡ Row 12 × correction factor 1,200 00 ÷ 1,424 31 = .842513
Source: Table 113

4. *Adjust for trend by successive subtraction of a correction factor from each chain relative.* This correction factor is $\frac{1}{12}$ of the amount by which the second January chain relative exceeds the first; $11.89 \div 12 = .9908$, in this case. This correction is seen in rows 11 and 12 of Table 114.[7]
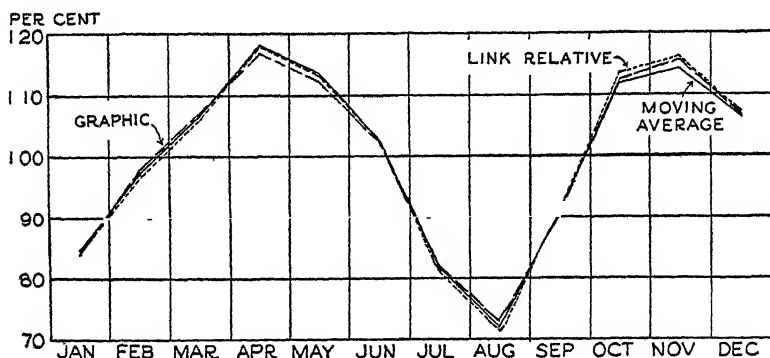
Chart 182. Seasonal Indexes of United States Magazine Advertising by three Methods, 1922-1929. (Data of Table 115.)

5. *Adjust the chain relatives to total 1,200.0.* The total of row 12 (excluding the second January figure) is 1,424.31. The correction factor is $1,200.00 \div 1,424.31 = .842513$, and each item in row 12 is multiplied by this number. The results, shown in row 13, total 1,200.1 and are the seasonal index numbers.

---

[7] This method of adjustment for trend seems appropriate for these data, since the trend is a straight line If the trend is an exponential curve, a logarithmic adjustment should be made. Beginning with step 3 the procedure is as follows:

3 *Obtain logarithms of the average link relatives and total them.* Since the logarithm of 100 is 2 0, the sum of the logarithms would tend to equal 24.000 were there no trend in the data (This will be clear if it is realized that successive multiplication of the twelve link relatives, taken as decimals, would obtain a product of 1.0 under the same circumstances)

4. *Adjust for trend by subtracting a correction factor from each logarithm* The correction factor is $\frac{1}{12}$ of the amount by which the sum of the logarithms exceeds 24

5. *Chain the logarithms of the corrected link relatives to the preceding month by successive addition.* This is the logarithmic equivalent of successive multiplication of the link relatives. The January logarithm is arbitrarily taken as 2 0, the logarithm of 100. The final logarithm in the chain is not the sum of the other numbers in that row, but is the sum of the December logarithmic chain relative and the January adjusted logarithmic link relative.

6. *Obtain the anti-logarithms of the logarithmic chain relatives* These are, of course, the chain relatives; that is, the seasonal variation of each month relative to January as 100. In looking up the anti-logarithms, the characteristics of the logarithms are taken as 2 or 1. This is because, as a matter of convenience, the link relatives are originally recorded as percentages rather than as decimals. Had decimals been used, the characteristics would each have been either 0 or -1.

7. *Adjust the chain relatives to total 1,200.0.* This is done in the usual fashion by multiplying by a correction factor. The results are the seasonal index numbers.

The link relative method of seasonal index construction has not so much to commend it as has the moving average method. The link relatives averaged together contain both trend and cyclical movements. Although the trend is subsequently removed, the process is effective only if the growth is one of constant amount or constant rate. Nor is the method so readily adaptable as the other to the construction of some of the more complex types of seasonal movements that will be described in the following chapter. Furthermore, it is a confusing method for most beginners. Yet it has some theoretical and practical advantages. It is a characteristic of time series that values of the original data for a given month are in part dependent on values existing during a few of the more recent preceding

TABLE 115

SEASONAL INDEXES OF UNITED STATES MAGAZINE ADVERTIS-
ING, 1922–1929, AS OBTAINED BY THREE METHODS

Month	Moving average method	Graphic method	Link relative method
January	84 8	84.0	84 3
February	97.2	97.9	96.6
March	106.6	106 9	106.0
April	118 2	116.9	118 0
May	113 5	112.4	113 2
June	102.6	102.4	102.8
July	81 6	81 9	81.1
August	72 0	72.7	71.2
September	91 2	90 4	91 0
October	111.9	112 7	113 5
November	114.1	115 7	115 8
December	106.3	106 4	106.6

Source· Tables 111, 112, and 114

months. Or, to put it another way, irregular movements are frequently of more than one month's duration. If then, let us say, March, April, and May are each unusually high months, nevertheless, April will not be high relative to March, or May high relative to April. Thus the link relatives in such an instance are less disturbed by an irregular movement than would be percentages of a 12-month moving average. Also, with the link relative method, cyclical peaks and troughs do not influence the percentages from which the index is computed to the same degree as with the moving average method. Another point in favor of the method is that it more completely utilizes the data. There is only one less link relative than the number of months available, whereas a 12-month moving average

(see Table 109) cuts off 6 months at each end. For short periods, therefore, the link relative method is very useful.

Comparison of results. The seasonal indexes by the last three methods, which employ a fairly refined technique, are given in Table 115. There is little to choose between the results.[8] As shown by Chart 182, the three lines are very difficult to distinguish from one another.

Adjustment for Seasonal

A seasonal index may be computed for the purpose of studying the seasonal movement itself—possibly in order to avoid the consequences of it, possibly in order to smooth out the seasonal fluctuations. On the other
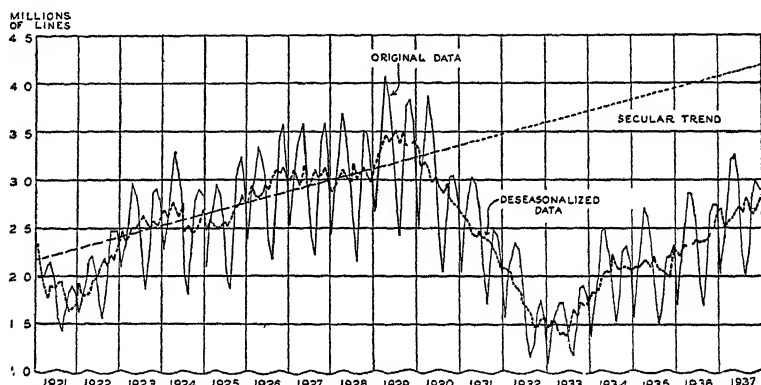
Chart 183. United States Magazine Advertising, Deseasonalized Data and Trend, 1921-1937. (Data of Table 116. For trend, see Table 89.)

hand, it may be that the computation of seasonal is but one step in the isolation of cyclical movements, either alone or in combination with secular trend. Often we are interested in studying the cyclical movements, and in this case it is necessary to adjust the original data, in turn, for seasonal movements and for trend. Sometimes, however, it may be desirable to study the combined effect of trend and cyclical movements. Thus business men, in making decisions, may consider not so much whether their sales are increasing relative to some vague trend as whether their sales are increasing or decreasing more than could be expected after taking the season of the year into consideration.

The mechanics of eliminating seasonal is to divide the original data by

---

[8] The three indexes are not strictly comparable as to period of time covered. The moving average method employs percentages of moving average from 1922 through 1929. The link relatives are for the same period. The graphic method, however includes readings for 1921 as well

TABLE 116

ELIMINATION OF SEASONAL VARIATIONS FROM UNITED STATES MAGAZINE ADVERTISING, 1921–1937

(Seasonal indexes were computed by 12-month moving average method)

Year and month (1)	Original data (2)	Seasonal index (3)	Deseasonalized data [Col 2 ÷ Col. 3] (4)
1921			
January .	1,979	84 8	2,334
February . .	1,981	97 2	2,038
March	2,005	106 6	1,881
April . . .	2,099	118.2	1,776
May ..	2,145	113 5	1,890
June	1,933	102 6	1,884
July..... . . .	1,573	81 6	1,928
August	1,402	72.0	1,947
September	1,620	91 2	1,776
October	1,824	111 9	1,630
November	1,903	114.1	1,668
December	1,807	106.3	1,700
1922			
January............	1,632	84.8	1,925
February.	1,768	97.2	1,819
March	1,922	106 6	1,803
April	2,171	118 2	1,837
May	2,215	113 5	1,952
June	2,046	102 6	1,994
July	1,705	81 6	2,089
August	1,566	72 0	2,175
September	1,940	91.2	2,127
October	2,470	111.9	2,207
November..	2,466	114 1	2,161
December....	2,464	106.3	2,318
1923			
January..	2,093	84.8	2,468
February	2,301	97 2	2,367
March	2,557	106 6	2,399
April...	2,963	118 2	2,507
May	2,850	113.5	2,511
June.	2,632	102.6	2,565
July	2,150	81.6	2,635
August	1,864	72.0	2,539
September...	2,277	91.2	2,497
October	2,873	111.9	2,567
November	2,900	114.1	2,542
December.... ..	2,773	106.3	2,609
1924			
January	2,281	84.8	2,690
February	2,557	97.2	2,631
March.	2,915	106.6	2,735
April	3,273	118.2	2,769
May.....	2,988	113.5	2,633
June	2,830	102.6	2,758
July	2,017	81.6	2,472
August	1,819	72.0	2,526
September	2,242	91.2	2,458
October	2,779	111.9	2,483
November .. .	2,900	114.1	2,542
December . ..	2,841	106.3	2,673

493

TABLE 116 (Continued)
ELIMINATION OF SEASONAL VARIATIONS FROM UNITED STATES MAGAZINE ADVERTISING
1921–1937

(Seasonal indexes were computed by 12-month moving average method)

Year and month (1)	Original data (2)	Seasonal index (3)	Deseasonalized data [Col 2 ÷ Col 3] (4)
1925			
January . .	2,111	84.8	2,489
February .	2,513	97.2	2,585
March	2,712	106.6	2,544
April	2,951	118.2	2,497
May	2,854	113.5	2,515
June	2,635	102.6	2,568
July	2,068	81.6	2,534
August	1,878	72.0	2,608
September	2,485	91.2	2,725
October	3,062	111.9	2,736
November	3,244	114.1	2,843
December......	2,960	106.3	2,785
1926			
January	2,385	84.8	2,812
February.. ...	2,854	97.2	2,936
March	3,030	106.6	2,842
April.. ..	3,343	118.2	2,828
May	3,236	113 5	2,851
June : .	3,024	102 6	2,947
July	2,368	81 6	2,902
August . ..	2,181	72 0	3,029
September ..	2,831	91 2	3,104
October	3,444	111 9	3,078
November	3,586	114.1	3,143
December	3,209	106 3	3,019
1927			
January	2,536	84.8	2,991
February.	3,005	97 2	3,092
March	3,255	106 6	3,053
April.....	3,497	118 2	2,959
May	3,577	113.5	3,152
June	3,012	102 6	2,936
July	2,420	296 6	2,939
August	2,229	72 0	3,096
September.	2,762	91.2	3,029
October.........	3,411	111.9	3,048
November	3,573	114.1	3,131
December...... . ..	3,176	106 3	2,988
1928			
January	2,447	84.8	2,886
February..	2,846	97.2	2,928
March......	3,212	106 6	3,013
April.......	3,675	118 2	3,109
May.	3,435	113 5	3,026
June..	3,061	102 6	2,983
July............. ..	2,583	81 6	3,165
August.......	2,158	72.0	2,997
September..	2,805	91 2	3,076
October..	3,499	111 9	3,127
November.	3,486	114 1	3,055
December.	3,172	106 3	2,984

TABLE 116 (Continued)
ELIMINATION OF SEASONAL VARIATIONS FROM UNITED STATES MAGAZINE ADVERTISING,
1921–1937

(Seasonal indexes were computed by 12-month moving average method)

Year and month (1)	Original data (2)	Seasonal index (3)	Deseasonalized data [Col. 2 − Col 3] (4)
1929			
January	2,684	84 8	3,165
February.	3,158	97.2	3,249
March...... . . .	3,601	106 6	3,378
April	4,082	118.2	3,453
May	3,875	113 5	3,414
June	3,547	102 6	3,457
July	2,864	81 6	3,510
August . .	2,430	72 0	3,375
September	3,162	91 2	3,467
October . .	3,760	111.9	3,360
November .. .	3,828	114 1	3,355
December	3,615	106 3	3,401
1930			
January . . .	2,505	75 1	3,336
February	3,024	96 2	3,143
March	3,416	107 5	3,178
April	3,877	123.6	3,137
May.	3,639	122 0	2,983
June.	3,354	111 1	3,019
July	2,451	83 3	2,942
August	2,057	71 7	2,869
September	2,598	87 7	2,962
October	3,021	107.9	2,800
November	3,042	110.5	2,753
December	2,820	103.3	2,730
1931			
January	2,001	75.1	2,664
February	2,539	96.2	2,639
March	2,762	107 5	2,569
April	3,026	123 6	2,448
May	2,971	122 0	2,435
)une	2,732	111.1	2,459
July	1,998	83.3	2,399
August	1,713	71.7	2,389
September..	2,069	87.7	2,359
October	2,480	107 9	2,298
November.	2,444	110.5	2,212
December..	2,170	103.3	2,101
1932			
January	1,570	75.1	2,091
February	2,000	96.2	2,079
March	2,184	107.5	2,032
April	2,348	123.6	1,900
May	2,278	122 0	1,867
June	1,903	111 1	1,713
July	1,394	83.3	1,673
August	1,173	71 7	1,636
September .	1,310	87.7	1,494
October	1,607	107.9	1,489
November	1,754	110 5	1,587
December	1,641	103 3	1,589

TABLE 116 (Continued)
ELIMINATION OF SEASONAL VARIATIONS FROM UNITED STATES MAGAZINE ADVERTISING.
1921–1937
(Seasonal indexes were computed by 12-month moving average method)

Year and month (1)	Original data (2)	Seasonal index (3)	Deseasonalized data [Col 2 ÷ Col 3] (4)
1933			
January . .	1,116	75 1	1,486
February . .	1,490	96 2	1,549
March	1,630	107 5	1,516
April ...	1,729	123 6	1,399
May . .	1,732	122 0	1,420
June . ..	1,544	111.1	1,390
July	1,272	83.3	1,527
August . .	1,184	71.7	1,651
September . .	1,407	87 7	1,604
October . . .	1,870	107.9	1,733
November . . .	1,899	110 5	1,719
December . .	1,791	103 3	1,734
1934			
January	1,375	75 1	1,831
February	1,765	96 2	1,835
March .. .	2,013	107 5	1,873
April...	2,469	123 6	1,998
May	2,501	122 0	2,050
June	2,271	111 1	2,044
July..... . .	1,853	83 3	2,224
August	1,534	71 7	2,139
September . . .	1,827	87.7	2,083
October	2,264	107.9	2,098
November	2,317	110.5	2,097
December...	2,136	103.3	2,068
1935			
January	1,581	75.1	2,105
February	2,014	96.2	2,094
March.....	2,276	107.5	2,117
April...	2,700	123 6	2,184
May	2,618	122.0	2,146
June	2,335	111.1	2,102
July.....	1,831	83 3	2,198
August...	1,497	71 7	2,088
September	1,812	87.7	2,066
October.....	2,181	107.9	2,021
November.	2,201	110 5	1,992
December........ ..	2,334	103 3	2,259
1936			
January.............	1,696	75 1	2,258
February.............	2,128	96 2	2,212
March........... ..	2,511	107.5	2,336
April.............. .	2,860	123.6	2,314
May................	2,852	122 0	2,338
June................ .	2,637	111.1	2,374
July.................	1,967	83.3	2,361
August..............	1,695	71.7	2,364
September............	2,084	87.7	2,376
October.............	2,637	107 9	2,444
November............	2,736	110.5	2,476
December............	2,731	103.3	2,644

TABLE 116 (Continued)

ELIMINATION OF SEASONAL VARIATIONS FROM UNITED STATES MAGAZINE ADVERTISING, 1921–1937

(Seasonal indexes were computed by 12-month moving average method)

Year and month (1)	Original data (2)	Seasonal index (3)	Deseasonalized data [Col. 2 ÷ Col. 3] (4)
1937			
January	2,031	75 1	2,704
February	2,399	96 2	2,494
March.....	2,762	107.5	2,569
April	3,206	123 6	2,594
May	3,258	122.0	2,670
June	3,023	111 1	2,721
July	2,235	83 3	2,683
August	2,018	71 7	2,815
September . . .	2,383	87 7	2,717
October	2,852	107 9	2.642
November	2,989	110.5	2,705
December .	2,893	103.3	2,801

Source. Tables 109 and 111, and data of Chart 177B

the seasonal index, as in Table 116. Note that two seasonal indexes have been used: one for the years before 1930, and a different one for 1930 through 1937, both of which were obtained by the 12-month moving average method. The results are shown in Chart 183. The deseasonalized data contain three elements: trend, cyclical movements, and irregular movements. Sometimes it is desirable at this stage of the analysis to smooth out the irregular variations from the deseasonalized data. A consideration of the technique for accomplishing this is reserved for Chapter XIX.

Test of seasonal. A comparison of the percentages of 12-month moving average with the seasonal indexes as plotted on Chart 175B indicates the closeness of agreement between the two. The same data are arranged differently in sections A and B of Chart 177. The closeness of the dots and circles to the seasonal index line in these charts constitutes a similar test of the adequacy of the seasonal index. Thus it is apparent that the fit in section A of Chart 177 is excellent, and much better than that in section B. It is also possible to measure the reliability of the different seasonal indexes, ascertaining the significance of their deviations from 100 per cent.[9] Again, it may be desired to test whether the seasonal index numbers used for the later period are significantly different from those used for the earlier period. Furthermore, it might be important to know whether the index number for a given month is significantly different from

---

[9] This is sometimes tested by means of analysis of variance (see Chapter XIII) and by means of the correlation ratio (see Chapter XXIII). These tests are subject to the limitations mentioned in this section.

that for some other month or from some other value for the same month. The application of the theory of sampling to seasonal index numbers involves theoretical difficulties, however, that have not been fully overcome. These difficulties arise because: (1) modified means, rather than means of all the data, are ordinarily used in constructing seasonal indexes; (2) the distributions from which these means are computed do not represent random distributions, since the irregular movements of a time series are not random occurrences.

A very practical test is to see whether the use of the seasonal index adopted appears to eliminate all of the seasonal movement. The data of Chart 183 have been rearranged in Chart 184 in a manner similar to Chart 176. Close inspection of this chart reveals a slight similarity in the pattern for the years 1921, 1922, 1923, and 1924. For instance, in each of these years February is lower than January. The patterns from 1925 through 1929 also exhibit a faint family resemblance. These facts indicate that the indexes adopted do not completely eliminate the seasonal movement. Perhaps it would have been better to have further subdivided the data into periods of time with a separate index for each; though if subdivided too finely, not enough years would be available in each subperiod to provide a reliable seasonal index. It might have been better to have constructed a moving seasonal, the procedure for which will be explained in the following chapter. It should not be concluded, however, that the seasonal indexes selected are poor generalizations. Though the indexes are not perfect, the different tests indicate that they are very satisfactory.

Selected References

E. C. Bratt: *Business Cycles and Forecasting,* Chapter II; Business Publications, Inc., Chicago, 1937. Mainly a consideration of economic factors
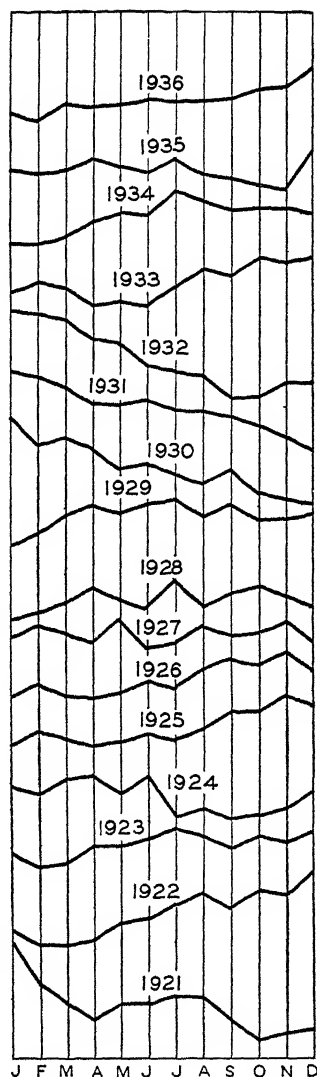
Chart 184. Deseasonalized United States Magazine Advertising Data, 1921-1936. (For purposes of comparison the different curves have been plotted close together on the same chart instead of on separate charts. Each curve is plotted to the same vertical scale, but at a different level. This arrangement is sometimes referred to as a multiple axis chart. Data of Table 116.)

R. W. Burgess: *Introduction to the Mathematics of Statistics*, Chapter IX; Houghton Mifflin Co , Boston, 1927.   Classifies and describes a number of methods of computing seasonal indexes.

R. E. Chaddock: *Principles and Methods of Statistics*, pages 337–353; Houghton Mifflin Co., Boston, 1925.

E. E. Day· *Statistical Analysis*, Chapter XVIII; Macmillan Co., New York, 1927. Explains the link-relative method in detail.

F. E. Croxton and D. J. Cowden· *Practical Business Statistics*, Chapter XIV; Prentice-Hall, Inc., New York, 1934.   Includes illustration of use of a short cut in the moving average method.

J. R. Stockton *An Introduction to Business Statistics*, Chapter IX; D. C Heath and Co., Boston, 1938.

CHAPTER XVIII

TYPES OF SEASONAL MOVEMENTS

In this chapter will be described methods of isolating some additional types of seasonal movements and further refinements of methodology. Of course, not all time series require the specialized treatment here outlined. The methods involved are somewhat more complex than those previously described, although the mathematics involved is entirely elementary.

Progressive Changes in Seasonal Pattern

Use of moving averages. The most satisfactory method of computing a seasonal index that is gradually changing in pattern, frequently called a *moving seasonal*, is based upon the fitting of curves to percentages of a moving average. Usually a 12-month moving average is used, the initial step therefore being the same as the most highly recommended method described in Chapter XVII. This will be the method used here.[1]

---

[1] A difficulty of a 12-month moving average was found in Chapter XVII to be that it smoothed out not only seasonal movements but part of the cycle as well. The percentages of 12-month moving average, therefore, included not only seasonal and irregular movements, but some of the cycle also. Furthermore, a 12-month moving average is not so smooth as might be desired. If extreme accuracy is desired, a much more laborious moving average is sometimes recommended. In Chapter XVI it was explained that weighted moving averages could be used for trends; and a curve, at the same time smoother and more flexible, could be obtained if the weight pattern selected were a smooth one. Macaulay recommends for the study of seasonal movements a considerably more complex weight system, involving 43 months, the diagram of which is shown herewith. It is stated that this moving average follows with almost complete accuracy *symmetrical* cyclical movements of from 30 to 120 months' duration. Actually, however,
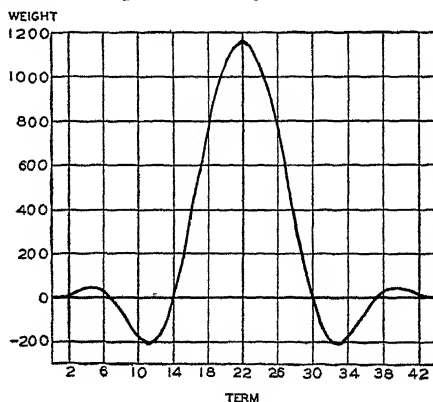
Weight Pattern of a 43-Term Moving Average. (Formula taken from Frederick R. Macaulay, *The Smoothing of Time Series*, p 148, National Bureau of Economic Research, New York, 1931.)

Computation of moving seasonal. As indicated above, the first step may be the computation of a 12-month moving average. Department store sales and moving average are plotted in Chart 185. The appearance of the chart indicates that, as compared with the seasonal fluctuations, the cyclical movements are relatively unimportant. A 12-month average is therefore sufficiently accurate for satisfactory results. The moving average has been extended freehand for 6 months in either direction to the edge of the chart, in order not to lose any of the original data. These extensions are shown by dotted lines. Next, the original data are divided by the moving average figures and recorded as percentages.

The method of deriving the moving seasonal is, from this point on, essentially graphic. On a large piece of graph paper which has been divided into twelve sections, one for each month, are plotted the percentage data. Thus, in the January rectangle of Chart 186, the January percentages *for each year* are plotted as shown by the thin line. In the first six sections, 1919 is connected with 1920 by a broken line to remind the computer that the 1919 figure is based on estimated data. Similarly, in the last six sections, 1935 is connected with 1936 by a broken line for the same reason. Also, there is a broken line connecting 1933 with 1932 and 1934 for January, February, and March. (This may be seen clearly in only the March section.) This line is broken because the percentage of moving average figure for those months was raised in an attempt to minimize the effect of the bank holidays during those months.

Next, each of the twelve curves of Chart 186 is smoothed by means of a 5-term moving average. In this chart the moving average data are represented by crosses. The object of the moving average is merely to aid in the location of the first approximation line, which was obtained by inspection and is shown by means of large dots. (Instead of freehand approximations, curves may be fitted mathematically to the percentages of moving average.) These dots are extended beyond the data as smoothed by the moving average, so as to include each year from 1919 through 1937. This extension provides us with a forecast for 1937 to be used in deseasonalizing that year's data. When drawing in this line, the statistician should

---

cyclical movements are not symmetrical and the use of this method may lead to unreasonable results. For further discussion, see Frederick R. Macaulay, *The Smoothing of Time Series*, National Bureau of Economic Research, New York, 1931. On page 148 is given the weight pattern. See page 159 for its degree of conformity to a sine curve. The formula for computing this moving average is given on page 73: "Take a 5-months moving total of a 5-months moving total of an 8-months moving total of a 12-months moving total of the data. To the results apply the following extremely simple weights: 7, −10, 0, 0, 0, 0, 0, 0, +10, 0, 0, 0, 0, 0, 0, −10, +7. Divide the final results by 9600." In Chapter VII Macaulay gives criteria for judging weight formulae, and in Chapter IV discusses a number of individual formulae, some of which he finds satisfactory

PER CENT
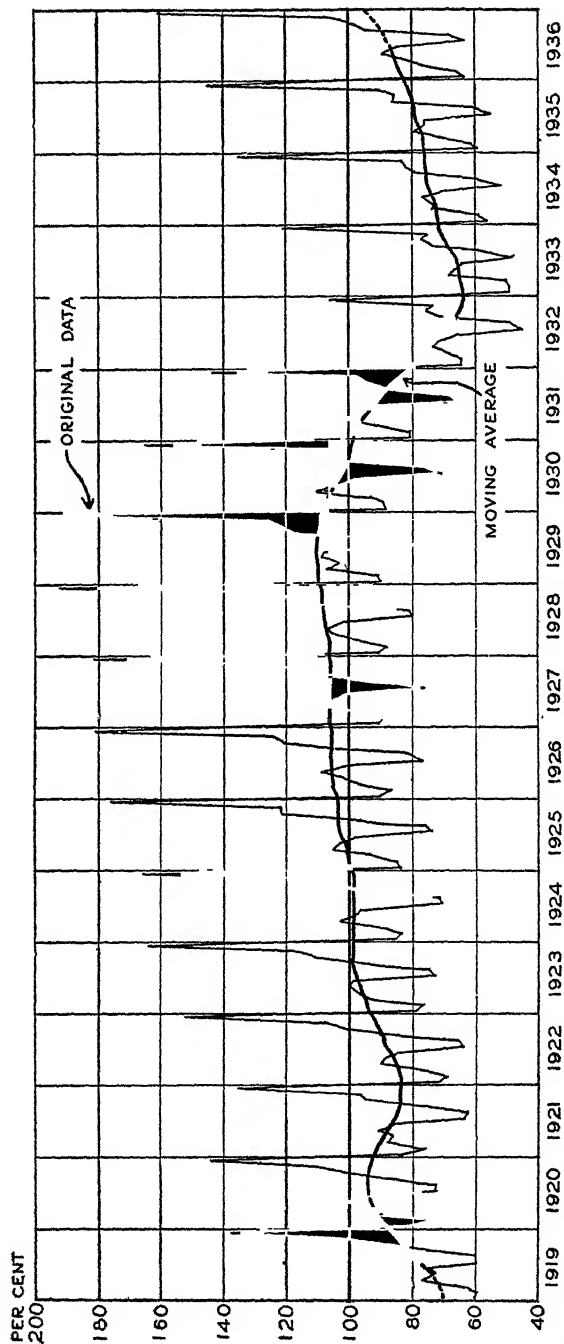
Chart 185. Federal Reserve Index of Department Store Sales, 1919-1936, and 12-Month Moving Average. (For source of data see Chart 168.)

TABLE 117

FIRST APPROXIMATION TO MOVING SEASONAL OF DEPARTMENT STORE SALES, 1919–1937

Month	1919	1920	1921	1922	1923	1924	1925	1926	1927	1928	1929	1930	1931	1932	1933	1934	1935	1936	1937
Jan	88.0	88.0	87.5	87.0	86.5	86.0	85.5	85.0	84.5	84.0	83.0	82.0	81.0	80.0	79.0	78.5	78.0	78.0	78.0
Feb	83.0	83.0	83.5	83.5	83.5	83.5	83.5	83.5	83.5	83.5	83.5	83.0	82.5	82.0	81.5	81.0	80.5	80.0	79.5
Mar	94.0	94.0	91.0	93.5	93.0	92.0	91.5	91.5	91.5	91.5	91.5	91.5	91.5	91.5	91.5	91.5	91.5	91.5	91.5
Apr	101.5	101.5	101.5	101.5	101.5	101.5	101.5	100.5	100.0	99.5	99.5	100.0	101.0	101.5	101.5	101.5	101.5	101.5	101.5
May	103.5	103.5	103.5	103.0	103.5	103.0	102.0	101.0	100.0	100.0	100.0	100.0	100.5	101.0	101.5	101.5	101.5	101.5	101.5
June	100.5	100.5	100.0	99.5	99.0	98.0	97.0	96.0	96.0	96.0	96.0	96.0	96.0	96.0	96.0	96.0	96.0	96.0	96.0
July	74.5	75.0	73.5	73.0	72.5	72.0	72.0	72.0	72.0	72.0	71.5	71.0	70.5	70.0	69.5	70.0	70.0	70.0	70.0
Aug	75.0	75.0	75.0	75.0	75.0	75.0	75.0	75.5	76.0	76.5	77.0	77.5	78.0	78.5	78.5	78.5	78.5	78.5	78.5
Sep	93.0	93.0	93.0	93.5	94.5	95.5	96.5	98.0	99.5	101.0	102.5	103.5	104.5	105.0	105.5	105.5	105.5	105.5	105.5
Oct	109.0	109.5	110.0	110.5	111.0	111.0	111.0	111.0	111.0	110.0	110.0	110.0	112.0	110.5	110.0	110.0	109.5	109.0	108.5
Nov	117.5	117.5	117.5	117.5	117.5	117.5	117.5	117.0	116.0	115.0	114.0	113.0	112.5	112.0	111.5	110.0	111.5	111.0	111.0
Dec	156.0	158.0	160.5	163.0	165.0	167.0	169.0	170.0	171.0	171.5	171.5	171.5	171.5	171.5	171.5	171.5	171.5	171.5	171.5
Total	1195.5	1197.5	1199.5	1201.0	1202.5	1202.0	1202.0	1201.0	1201.0	1201.5	1200.5	1200.0	1200.5	1200.0	1198.5	1196.5	1195.0	1194.0	1193.0

TABLE 118

MOVING SEASONAL INDEX OF DEPARTMENT STORE SALES, 1919–1937

Month	1919	1920	1921	1922	1923	1924	1925	1926	1927	1928	1929	1930	1931	1932	1933	1934	1935	1936	1937
Jan	88.5	88.0	87.5	87.0	86.5	86.0	85.5	85.0	84.5	84.0	83.0	82.0	81.0	80.0	79.0	78.5	78.0	77.5	77.0
Feb	83.5	83.5	83.5	83.5	83.5	83.5	83.5	83.5	83.5	83.5	83.5	83.0	82.5	82.0	81.5	81.0	80.5	80.0	79.5
Mar	95.0	91.5	91.5	93.5	93.0	92.5	92.0	91.5	91.5	91.5	91.5	91.5	91.5	91.5	92.0	92.5	93.0	93.5	94.0
Apr	101.5	101.5	101.5	101.5	101.5	101.5	101.5	101.0	100.0	99.5	99.5	100.5	101.0	101.5	101.5	101.5	101.5	101.5	101.5
May	103.5	103.5	103.5	103.0	102.0	101.5	101.0	100.5	100.0	99.5	99.0	100.0	100.5	101.0	101.5	101.5	101.5	101.5	101.5
June	101.5	101.0	100.0	99.0	98.0	97.0	96.5	96.0	95.5	95.5	96.0	96.0	96.0	96.0	96.0	96.0	96.0	96.0	96.0
July	76.0	75.0	74.0	73.0	72.5	72.0	71.5	71.5	71.0	71.5	71.0	70.5	70.0	69.5	69.5	69.5	70.0	70.0	71.0
Aug	75.0	75.0	75.0	75.0	76.0	75.0	75.0	75.5	76.0	76.5	77.0	77.5	78.0	78.5	79.0	79.0	79.0	79.0	79.0
Sep	93.0	93.0	93.0	93.5	94.5	95.5	96.5	98.0	99.5	101.0	102.5	103.5	104.5	105.0	106.0	106.5	106.5	106.5	106.5
Oct	109.0	109.5	110.5	110.5	111.0	111.0	111.0	110.5	111.0	110.0	110.0	110.0	110.5	110.5	110.0	109.5	109.0	108.5	108.0
Nov	117.5	117.5	117.5	117.5	117.5	117.5	117.0	116.5	116.0	115.0	114.0	113.0	112.5	112.0	115.5	115.0	115.5	115.5	115.0
Dec	156.0	158.0	162.5	163.0	165.0	167.0	169.0	170.0	171.0	171.5	171.5	171.5	171.5	172.0	172.5	173.0	173.5	174.0	174.5
Total	1200.0	1200.0	1200.0	1200.0	1200.0	1200.0	1200.0	1200.0	1200.0	1200.0	1200.0	1200.0	1200.0	1200.0	1200.0	1200.0	1200.0	1200.0	1200.0

503
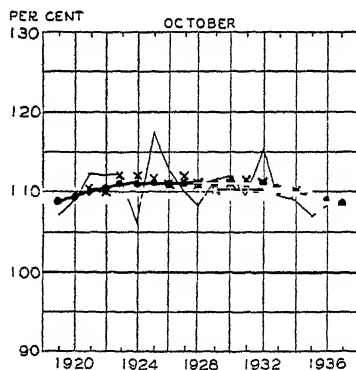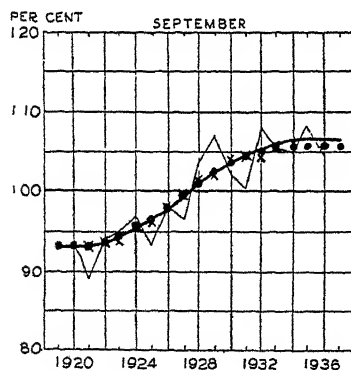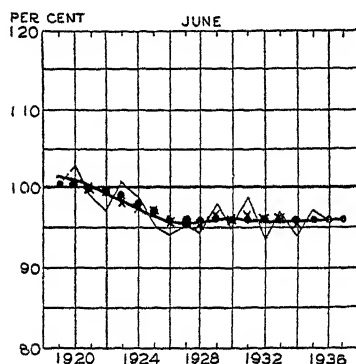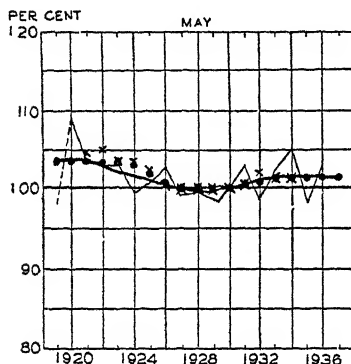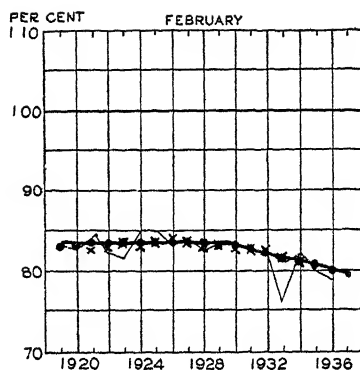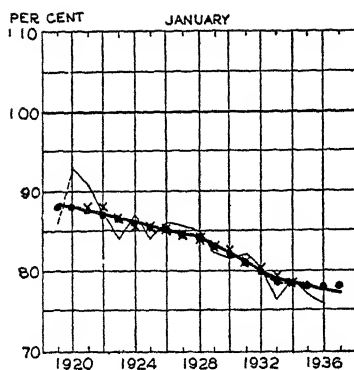
Chart 186A. Graphic Determination of Moving Seasonal of Department Store Sales 1919-1937. (For source of data see Chart 168.)
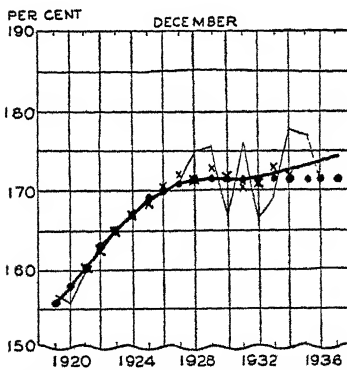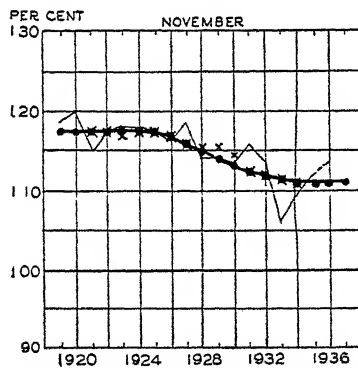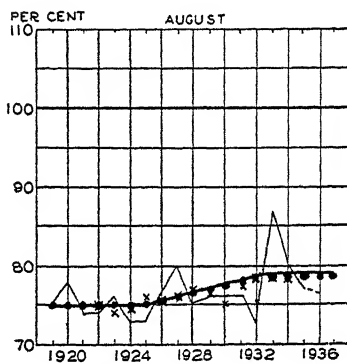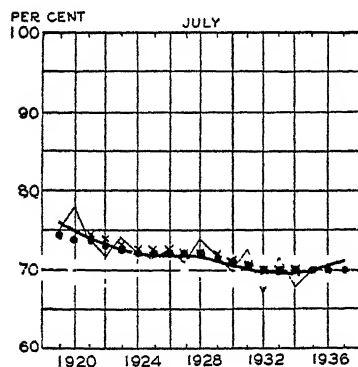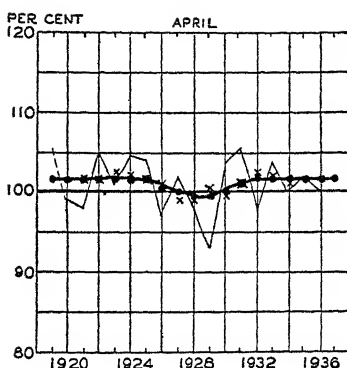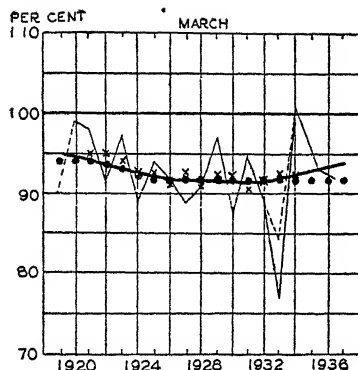
Chart 186B.   Graphic Determination of Moving Seasonal of Department Store Sales.
1919–1937 (continued).

505

draw relatively simple curves, and also attempt to be fairly conservative at the end years. "Conservatism" means here that the curves should, wherever it seems at all reasonable, have a tendency to flatten out rather than incline more steeply at the ends. For it must be remembered that these trends are not affected by the same factors associated with secular trend; they will not continue in a given direction indefinitely, but are more likely to move to a certain level and remain stabilized until new factors bring about a change in one direction or another. An additional reason for conservatism is that the first and last 6 months are at least partly estimated and not entirely trustworthy.

Annual values for each month are now read from the dots and entered in a table such as Table 117. The figures in the columns are not the final seasonal indexes, for each year does not total 1200 per cent. The statistician must now go back to his chart, and draw new lines, keeping three purposes in view: (1) make each year total 1200; (2) keep the curve in each rectangle smooth; (3) obtain final curves which fit the data. This is, of course, a highly subjective procedure and should not be undertaken by one lacking experience or not familiar with the series. Final results are shown in Table 118 and by the heavy lines of Chart 186.

If the reader will now turn to Chart 187, he will see the moving seasonal pattern for this series as shown by the dotted line. The changes in the seasonal pattern are so gradual that one is tempted to think that they are insignificant. However, if Chart 188 is referred to, it is apparent that there is a considerable difference between the 1919 seasonal pattern and the projected pattern for 1937. The more recent year shows considerably more amplitude of fluctuation, but most noticeable is the great increase in the relative importance of the Christmas trade.

The procedure for computing a moving seasonal may be summarized briefly as follows:[2]

(1) Compute 12-month moving averages of original data.

(2) Divide original data by these moving average figures and express as percentages.

(3) Plot these percentages each month, by years.

(4) Smooth with a 5-term (or other) moving average.

(5) Draw a freehand trend line and read values from it.

(6) Adjust these first approximation figures to total 1200 for each year, at the same time retaining a smooth, well-fitting trend.

[2] This procedure is taken from "Use of Moving Averages in the Measurement of Seasonal Variations," by Aryness Joy and Woodlief Thomas, *Journal of the American Statistical Association*, Vol. XXIII, September 1928, pp. 247-249.
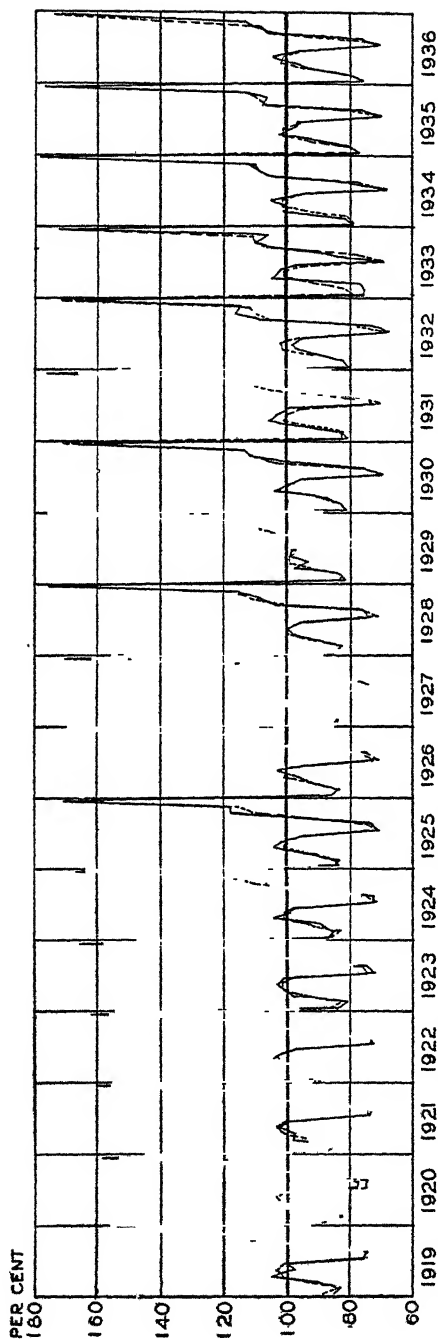
PER CENT

Chart 187. Percentages of 12-Month Moving Average and Moving Seasonal Pattern of Department Store Sales, 1919-1936. (This chart is a rearrangement of the data represented by the two solid lines of Chart 186.)

507

If a series that contains a moving seasonal is deseasonalized by a stable seasonal index, the adjusted data will contain not only the ordinary type of irregular movements but additional irregularities where the stable seasonal index has under-corrected for seasonal in some places and over-corrected in others. If the series has been adjusted by a moving seasonal, the resulting series should be somewhat smoother. That this seems to be the case with department store sales can be seen by a comparison of sections A and B of Chart 191. Of course, it is true that, if the moving
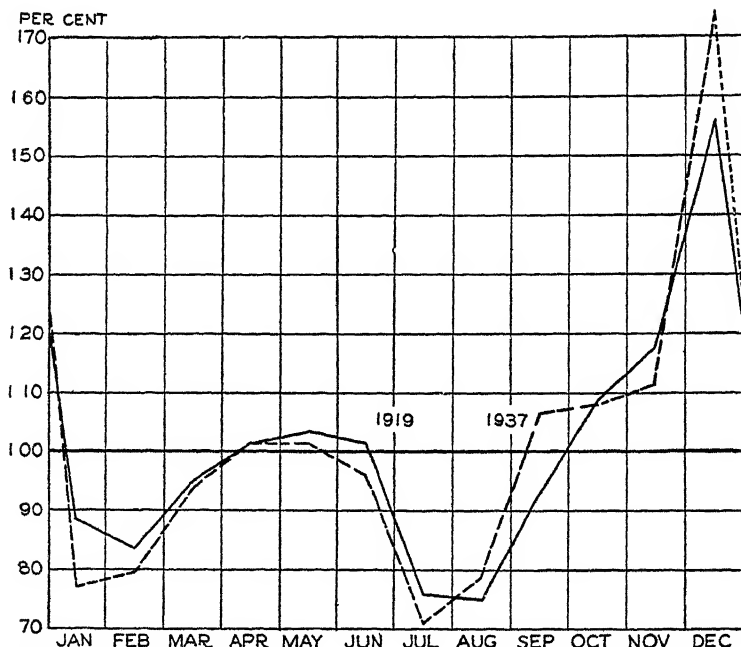
Chart 188. Seasonal Pattern of Department Store Sales for 1919 and 1937. (Data of Table 118.)

seasonal has been made too flexible, it may be merely a combination of a stable seasonal and smoothed random fluctuations, and deseasonalizing by such an index may remove not only the seasonal movements but some of the fluctuations of an irregular character as well. In the present instance the smoothness of the freehand curves of Chart 186 and the closeness of their fit to the percentages of moving average make this limitation seem unimportant. Furthermore, if the data are eventually to be partially smoothed of irregular movements, this preliminary smoothing probably does no great harm.

Sudden Variations in Seasonal Pattern

Seasonal patterns may change abruptly, rather than gradually, and then the device of a moving seasonal would be inapplicable. Such changes may involve merely the relative importance of two consecutive months, or may involve a change in the entire pattern. The most obvious change of the first type is that occasioned by the varying date of Easter, which may range from March 22 to April 25.

Adjustment for Easter. A number of industries are affected materially by Easter. Department store sales are one type of series which we should expect to be affected. A late Easter will tend to make April sales heavy relative to March, and, within limits, the later in April that Easter occurs, the greater is this tendency. On the other hand, when Easter occurs in March, March sales, and possibly February sales, will be increased.

A procedure for making the Easter adjustment is as follows:[3]

1. *Express the original March and April values each as percentages of 12-month moving average.* This has already been done if the seasonal has been computed by the method advocated in this book. If a weighted moving average has been computed—for instance, in computing a moving seasonal—that may be used instead of the 12-month average. These percentages are, of course, estimates of seasonal-irregular movements.

2. *Subtract the March seasonal index numbers from these March percentages, and the April index numbers from the April percentages.* If a moving seasonal has been computed, there will be a different number to subtract each year. See columns 4 and 7 of Table 119. These are called March and April residuals respectively, since they are variations remaining after taking the normal seasonal movement into account. They are due in part to irregular movements, but also to the fact that Easter sometimes occurs early and sometimes late.

3. *Subtract the March residuals from the April residuals as in column 8 of Table 119.* These differences between the April and March residuals are presumably due in large part to the varying date of Easter. Let us call these differences Easter residuals.

4. Next, it must be discovered whether or not these second residuals actually do vary in accordance with the date of Easter. In Chart 189 are plotted, by years, these Easter residuals and the date of Easter (data are from Table 120). It is apparent that there is a marked tendency for early Easter to increase March sales relative to April, and for late Easter

---

[3] This procedure is based on an article by Leroy M. Piser, "The Adjustment of Time Data for the Influence of Easter," *Journal of the American Statistical Association,* Vol XXIX, June 1934, pp. 190–191. Piser, however, does not include step 2 in his computations.

TABLE 119

COMPUTATION OF EASTER RESIDUALS FOR DEPARTMENT STORE SALES, 1919–1936

Year (1)	March			April			Easter residual [Col. 7 − Col. 4] (8)
	Per cent of moving average (2)	Seasonal index number (3)	Residual [Col. 2 − Col. 3] (4)	Per cent of moving average (5)	Seasonal index numbers (6)	Residual [Col. 5 − Col. 6] (7)	
1919	90 3*	95.0	−4.7	105.5*	101.5	4.0	8 7*
1920	99 1	94.5	4.6	99.1	101.5	−2 4	−7.0
1921	98.0	94.0	4.0	97.7	101 5	−3 8	−7.8
1922	90.6	93.5	−2.9	105.0	101.5	3.5	6.4
1923	97.5	93.0	4.5	100.9	101.5	−.6	−5 1
1924	89.1	92.5	−3.4	104.4	101 5	2 9	6.3
1925	93.7	92.0	1.7	104.0	101 5	3 5	1.8
1926	92.0	91.5	.5	96 6	101.0	−4.4	−4 9
1927	89.1	91.5	−2.4	102.3	100.0	2.3	4 7
1928	90.6	91.5	−.9	97 7	99.5	−2.2	−1 3
1929	97.2	91.5	5.7	93.3	99 5	−6.2	−11 9
1930	87.2	91.5	−4.3	104.2	100 5	3.7	8 0
1931	94.8	91.5	3.3	105.6	101.0	4.6	1 3
1932	90.3	91.5	−1.2	97.5	101.5	−4 0	−2 8
1933	84.0†	92.0	−8.0#	104.3	101.5	2 8	10.8†
1934	100.8	92.5	8.3	100.1	101.5	−1.4	9.7
1935	92.6	93.0	−.4	102.3	101.5	.8	1 2
1936	91.8	93.5	−1.7	100.2	101.5	−1.3	.4

* Based on freehand extensions of 12-month moving average
† Adjustment has been made for bank holiday   Number without adjustment was 76 9.
# −15 1 if measured from unadjusted percentages of moving average
‡ 17 4 if measured from unadjusted percentages of moving average.
Source. Data of Chart 187.

510

to stimulate April sales.   This tendency was less perfect during some of
the general depression years, especially 1932 and 1935.   But this chart
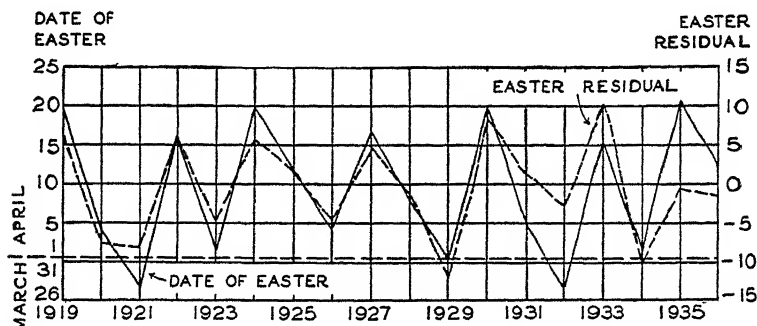
Chart 189.   Date of Easter and Easter Residuals for Department Store Sales, 1919–
1936.   (Data of Table 120)

does not tell us *how much, on the average, April sales are increased over
those of March for each additional day later that Easter occurs.   Such an
estimate can be obtained if the residuals are plotted, not by years, but with*

TABLE 120

DATE OF EASTER, EASTER RESIDUALS, AND EASTER CORRECTION
FACTOR, DEPARTMENT STORE SALES DATA, 1919–1936

Year	Date of Easter		Easter residuals
	Month	Day	
1919	April	20	8.7
1920	April	4	— 7.0
1921	March	27	— 7.8
1922	April	16	6.4
1923	April	1	— 5 1
1924	April	20	6.3
1925	April	12	1.8
1926	April	4	— 4.9
1927	April	17	4.7
1928	April	8	— 1.3
1929	March	31	—11.9
1930	April	20	8.0
1931	April	5	1.3
1932	March	27	— 2.8
1933	April	16	10.8
1934	April	1	— 9.7
1935	April	21	1.2
1936	April	12	.4

Source: Easter residuals from Table 119.   Date of Easter from New York
World Telegram, *The World Almanac and Book of Facts, 1934,* p. 66

Easter date along the horizontal axis and trend line fitted to the data as plotted. The slanting line of Chart 190, which was fitted by inspection, suggests that a change in one day in the date of Easter is responsible for an increase of about .75 per cent in April sales over those of March. The estimating line may be fitted mathematically if desired. However, it seems unrea-
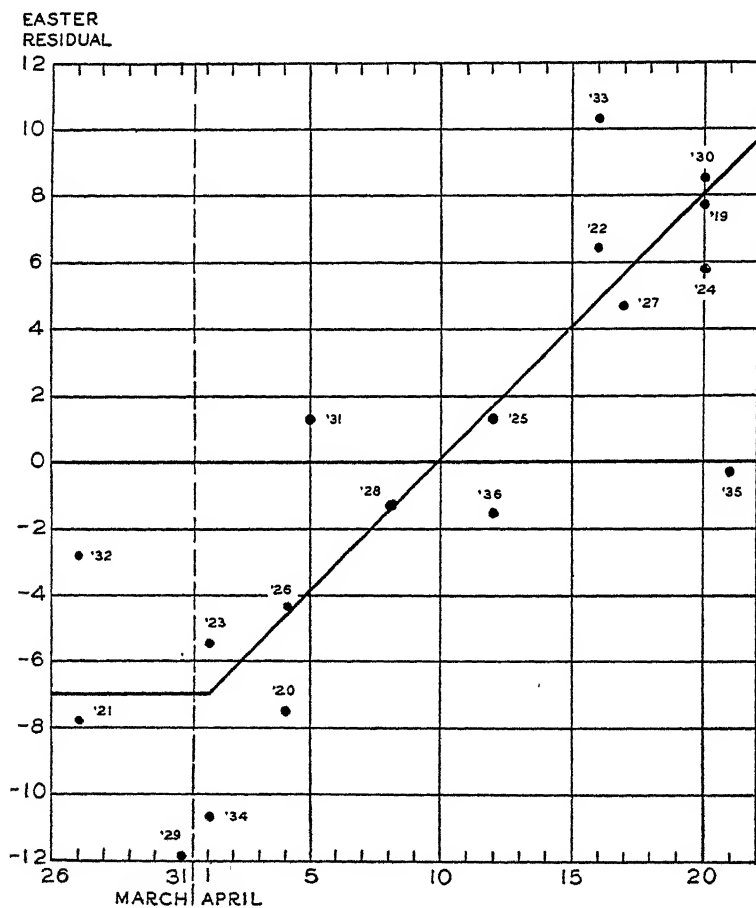
EASTER
RESIDUAL

Chart 190.   Graphic Estimation of Easter Correction Factor for Department Store Sales, 1919-1936.   (1933 is subjectively adjusted for closing of banks.   Data of Table 120.)

sonable to attach much weight to years such as 1933, for whose unusual behavior a definite reason can be assigned.

It is to be noted that the estimating line is drawn horizontally throughout March.   An Easter which occurs April 1 or earlier, has deprived April of any Easter sales whatever; no more damage can be done regardless of how

early in March Easter occurs. It is, of course, possible, if people shop very early, that an Easter occurring early in March may increase February sales relative to March. It is not often, however, that the statistician will consider it worth while to make that adjustment.

5. *Next, read off the correction factor for each date of Easter from the trend line.* The correction factor for all possible dates of Easter is given in column 2 of Table 121.

TABLE 121

EASTER CORRECTION FACTORS FOR DEPARTMENT STORE
SALES DATA

Date of Easter (1)	Gross correction factor (2)	Net correction factor applicable to each month* [Col 2 ÷ 2] (3)
March	−7.0	−3 5
April:		
1	−7.0	−3.5
2	−6 2	−3.1
3	−5 4	−2 7
4	−4.6	−2.3
5	−3 8	−1 9
6	−3.1	−1 6
7	−2.3	−1 2
8	−1.5	− 8
9	− .7	− .4
10	.1	0
11	.9	4
12	1.7	.8
13	2 5	1.2
14	3.3	1.6
15	4 1	2.0
16	4.8	2 4
17	5 6	2.8
18	6.4	3 2
19	7.2	3 6
20	8.0	4 0
21	8.8	4.4
22	9 6	4 8
23	10.4	5.2
24	11.2	5 6
25	11.9	6 0

* These values are to be added algebraically to April and sub-tracted algebraically from March
Source: Read from Chart 190

6. *Now, divide the correction factor by two,* since what April sales gain by a late Easter, March sales lose, and vice versa. See Table 121, column 3.

7. *Finally, add this amount algebraically to the April seasonal index num-
bers and subtract it algebraically from the March numbers, as in Table 122.*

TABLE 122

ADJUSTMENT OF MARCH AND APRIL SEASONAL INDEX NUMBERS FOR VARIATION IN
DATE OF EASTER, DEPARTMENT STORE SALES DATA, 1919–1937

Year	Date of Easter	Net correction factor*	March seasonal		April seasonal	
			Uncor- rected	Corrected [Col 4 − Col 3]	Uncor- rected	Corrected [Col 6 + Col 3]
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1919	April 20	4 0	95 0	91 0	101 5	105 5
1920	April 4	−2 3	94.5	96 8	101 5	99.2
1921	March 27	−3.5	94.0	97 5	101.5	98 0
1922	April 16	−2.4	93 5	95 9	101.5	99 1
1923	April 1	−3 5	93.0	96 5	101.5	98 0
1924	April 20	4 0	92 5	88.5	101 5	105 5
1925	April 12	8	92.0	91 2	101 5	102.3
1926	April 4	−2 3	91 5	93 8	101 0	98.7
1927	April 17	2.8	91.5	88.7	100 0	102 8
1928	April 8	⌣ 8	91.5	92.3	99.5	98.7
1929	March 31	−3 5	91.5	95.0	99 5	96 0
1930	April 20	4 0	91.5	87.5	100 5	104 5
1931	April 5	−1.9	91.5	93 4	101 0	99 1
1932	March 27	−3.5	91.5	95 0	101.5	98 0
1933	April 16	2.4	92 0	89 6	101 5	103.9
1934	April 1	−3.5	92 5	96 0	101 5	98 0
1935	April 21	4.4	93.0	88.6	101.5	105.9
1936	April 12	.8	93.5	92.7	101.5	102.3
1937	March 28	−3 5	94 0	97.5	101.5	98 0

* To be added algebraically to April and subtracted algebraically from March.
Source Tables 118 and 121.

Seasonal is now removed in the usual way, dividing the original data
by the revised seasonal index. If a series is deseasonalized by a seasonal
index that takes no account of the effect of the varying date of Easter,
the adjusted data will contain not only the ordinary type of irregular
movements, but additional irregularities which are due to the varying date
of Easter. If the seasonal index which is used takes account of the vary-
ing date of Easter, the deseasonalized data will therefore be smoother. It
is possible that accidental variations between March and April may mis-
takenly be attributed to variation in the date of Easter, and the correc-
tion for Easter may, in fact, become merely a roundabout method of
smoothing out such irregular movements. The closeness of the dots to
the slanting line of Chart 190 makes it seem highly improbable that this

qualification is of much importance in the case of department store sales. Bearing this qualification in mind, however, and the similar one mentioned on page 508 concerning the moving seasonal, it should be observed that the deseasonalized data shown in the three parts of Chart 191 become progressively smoother as adjustment is made by progressively more refined seasonal indexes.
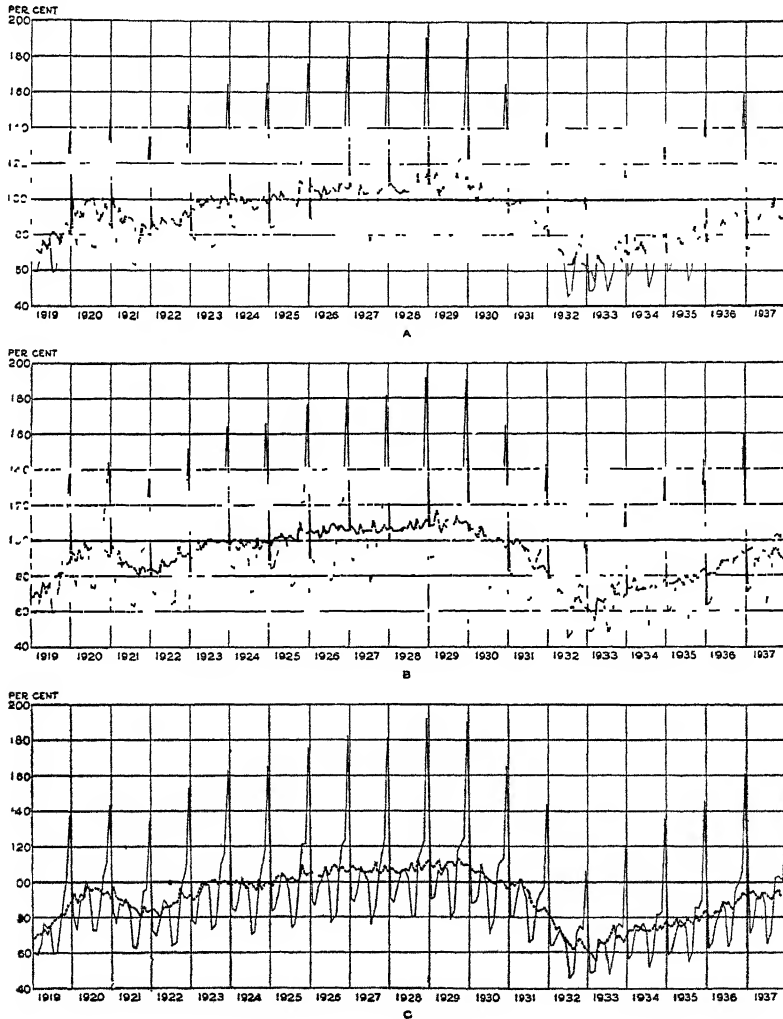
Chart 191.   Department Store Sales and Different Adjustments for Seasonal Variation, 1919-1937; A. Adjustment for Stable Seasonal, B. Adjustment for Moving Seasonal, C. Adjustment for Moving Seasonal and Varying Date of Easter.   (Computed from original data of Chart 185.)

Sudden changes in entire seasonal pattern.　In Chapter XVII our study of the seasonal behavior of United States magazine advertising led to the conclusion that there was a material change in the seasonal pattern after 1929. The remedy adopted was to break the whole period into two sub-periods: one for the years before 1930, and the other from 1930 on. Perhaps an even more clear-cut illustration is that of automobile production. Before 1935 it was customary to hold the New York automobile show in January. In 1935, however, a show was held in November (as well as in the preceding January); and since 1935 there has been but one show a year, that being in early November. The result has been two seasonal lows and two seasonal highs a year, instead of one. Whereas previously the low was in the fall and the high in the spring, a few months after the show, now there is one low just preceding and another just after the show, and a high coinciding closely with the show as well as one in the spring, about April. Consequently it seems advisable to compute two seasonal indexes rather than one. In part A of Chart 192 are shown the original data and the data as deseasonalized by a stable seasonal, while part B of the chart shows the data deseasonalized by two seasonal indexes, one running from April 1930 through March 1935, and the other from April 1935 through March 1937. The reason that a year from April through March, rather than a calendar year, was taken is that the automobile show did not change the fundamental pattern, which is a spring high and a fall low, but merely added another irregularity during the winter. By breaking the year between March and April, the problem of how to handle the calendar year 1935 was solved. Of course, the seasonal index for the last two 12-month periods is not very reliable, being an average of only two items　Further discussion of these indexes will be given on page 527.

Short-time shifts in timing.　The varying date of Easter affects materially only March and April, and the automobile show affects chiefly a few months preceding and following it. Weather conditions, however, which also vary from year to year, may result in early harvests one year and late harvests the next; and not only may the marketing of the product begin at different times in different years, but the flow of goods during the entire year may be affected, the effect being to shift the whole pattern a few months to the left or right. Likewise, consumer demand may vary in timing, depending on how early the weather changes.

Such shifting seasonal patterns present a difficult problem. Perhaps the most practical solution is to regard the problem as a special case of a sudden change in entire pattern, to group together the years (not necessarily adjacent) which show the same timing in their seasonal turns, and to compute as many seasonal indexes as there are groups of years. In computing such indexes, there is no reason why the calendar year must be
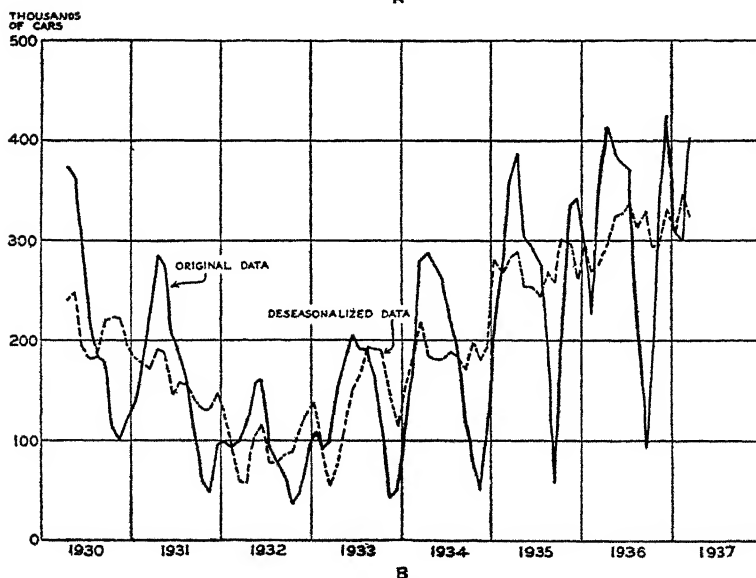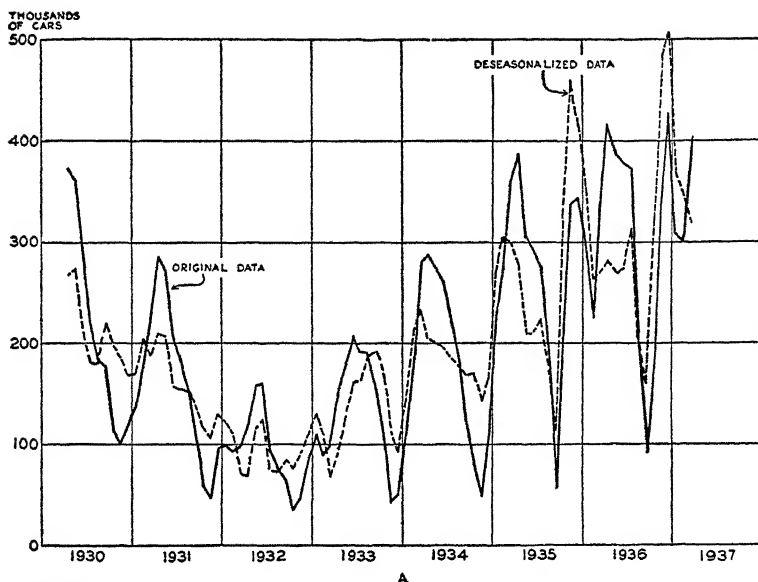
Chart 192. United States Passenger Car Production as Deseasonalized (A) by Single Index and (B) by Separate Indexes for Two Sub-periods, April 1930-March 1937. (Original data were compiled by the United States Bureau of the Census in Cooperation with the Automobile Manufacturers Association, and published in U. S. Department of Commerce, *Survey of Current Business, 1932 Supplement*, pp. 274–275, *1936 Annual Supplement*, p. 147, March 1937. p. 55.)

taken as a unit.   Rather, if the subject matter has to do with agriculture, the year should be related to the crop year.   Perhaps the central month should be the seasonal high or the seasonal low.

Varying amplitude.   Some economic series retain more or less the same general seasonal pattern from year to year, but have a tendency to vary rather suddenly in amplitude.   This is particularly true of stocks of agricultural commodities.   For example, stocks of agricultural crops show varying seasonal amplitude from year to year depending upon the amount carried over from the preceding year, the size of the harvest, and the amount consumed.   Likewise, shipments of livestock are likely to vary in the amplitude of their seasonal swing.   Here the variation may have something to do with the advantage of immediately selling the livestock, as compared with holding them for further fattening or a price increase. Since the relative advantages of these policies, as explained on page 156, is likely to vary in cycles, so the amplitude of the seasonal variation is likely to change in cycles, and the change in pattern might conceivably be treated as a moving seasonal.   Another borderline case is that of increased seasonal amplitude in manufacturing, brought about by a general cyclical tendency toward hand-to-mouth buying.   It is apparent that this change also might be thought of as a moving seasonal, the progression being cyclical rather than trend-like.

It must be apparent that, when the amplitude is not changing gradually but changing suddenly, and in the main unpredictably, a moving seasonal cannot overcome this problem any better than it can that of short-time shifts in entire pattern or in timing.   Any of the types of seasonal hitherto described would in some years over-correct the data and in other years under-correct it.   The object of the following procedure then is: (1) to discover for each year whether the seasonal amplitude is greater or less than average (more specifically, to measure how much the actual seasonal varies with a given change in the seasonal index); (2) to adjust the seasonal index so that it will have the correct amplitude.   The procedure[4] is somewhat analogous to the Easter adjustment.   Receipts of sheep and lambs at primary markets are taken as an illustration.

1.   *Express the seasonal index for each year as deviations from 100.*   See column 3, Table 123.

2.   *Express the original data as percentages of 12-month moving average.* A weighted moving average may be substituted for the 12-month moving average if preferred.   In the present illustration a 12-month moving aver-

---

[4] This procedure is based upon an article by Simon S. Kuznets, "Seasonal Pattern and Seasonal Amplitude: Measurement of their Short-Term Variations," *Journal of the American Statistical Association*, Vol. XXVII. March 1932. pp. 9-20.

age was smoothed slightly by inspection.   The percentages are in column 4
of Table 123.

TABLE 123

Seasonal Index and Percentage Deviations from 12-month Moving Average, Adjusted to Average Zero, for Receipts of Sheep and Lambs, 1934

Month	Seasonal index		Percentage of 12-month moving average		
	Percentage	Percentage deviation [Col 2 − 100]	Uncorrected percentage	Corrected percentage [Col 4 × c][k]	Percentage deviation [Col 5 − 100]
(1)	(2)	(3)	(4)	(5)	(6)
January .	84	−16	85.2	86	−14
February	75	−25	68 0	68	−32
March .	80	−20	73 0	74	−26
April	93	− 7	85 1	86	−14
May .	98	− 2	97 4	98	− 2
June.	92	− 8	82 8	83	−17
July	94	− 6	97 8	98	− 2
August	119	19	118.6	119	19
September	143	43	149 7	151	51
October .	150	50	182 3	184	84
November	95	− 5	82 4	83	−17
December .	76	−24	69 5	70	−30
Total	1199	− 1	1191 8	1200	0
Average	100	0		100	0

* $c = 1200 0 ÷ 1191 8 = 1 00688$
Source Original data from United States Department of Commerce, *Survey of Current Business*, 1932 Annual Supplement, p 165; 1936 Supplement, p 96, March 1937, p 43.

3. *Adjust these percentages so that they will total (and therefore average) zero, algebraically, for each year.* This may be done by averaging algebraically the percentages for each year, and then subtracting these averages from each of the items for the corresponding year; or the percentages may be multiplied in the usual fashion by a correction factor so that they will total 1200, and then 100 may be subtracted from each corrected percentage. The latter procedure is followed here, and is shown in columns 5 and 6 of Table 123.

4. A comparison of these two sets of data will tell us which has the greater amplitude, the seasonal index (column 3) or the percentages of moving average (column 6). From Chart 193 it is apparent that in 1934 the latter varied more, while in 1932 the seasonal index had the greater amplitude. The 1934 seasonal index would, therefore, fail to remove all of the seasonal, while in 1932 the seasonal index would over-correct the data. (Incidentally it might be noted that, except for amplitude, the variations conform rather closely to the seasonal index in both of these years.)
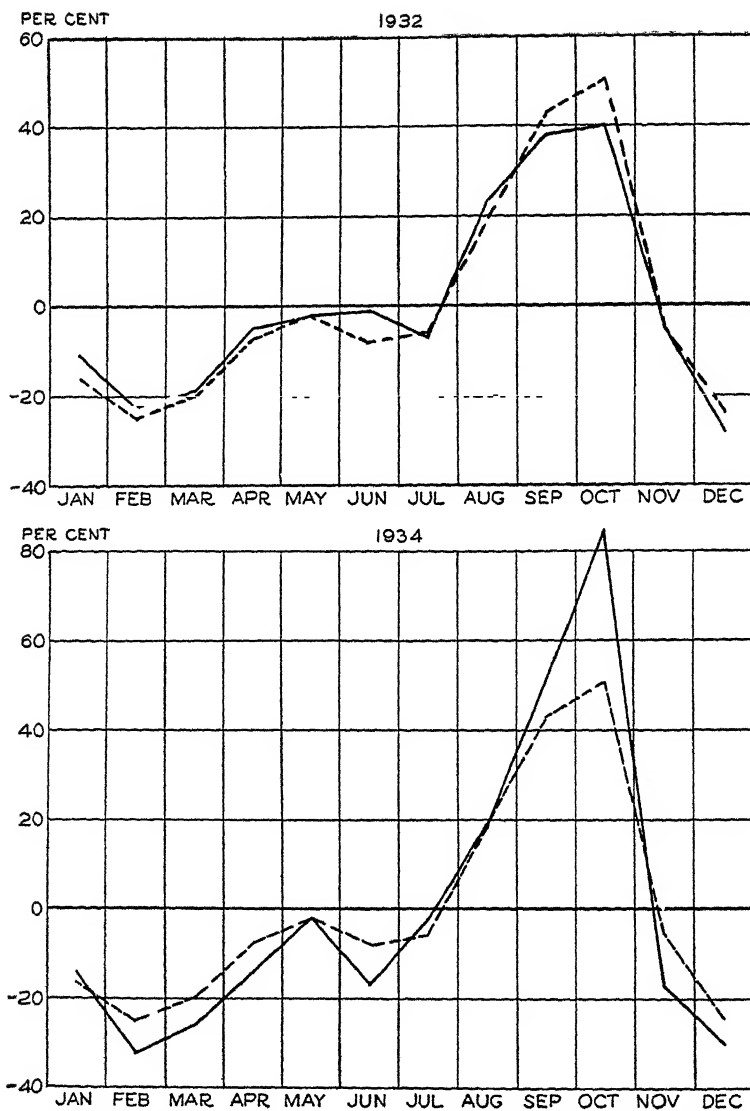
Chart 193.   Stable Seasonal Index and Percentage Deviations from 12-Month Moving Average of Receipts of Sheep and Lambs, 1932 and 1934.   (Stable seasonal is indicated by broken line; percentage deviations by solid line.   For source of original data see Table 123.   1934 data are from Table 123.)

A rearrangement of the material of Chart 193 will yield more useful information. In this chart the deviations were arranged chronologically (by months) and connected by straight lines. If the horizontal scale is used to represent the values of the seasonal index, or x, and the vertical axis to represent deviations from moving average, or y, as in Chart 194, we can *estimate about how much, on the average, the percentage deviations vary as the seasonal index numbers vary.* The broken diagonal straight line of the 1934 chart informs us that, for every increase of 1 per cent in x (the seasonal index), the y values change (in the same direction) about 1.4
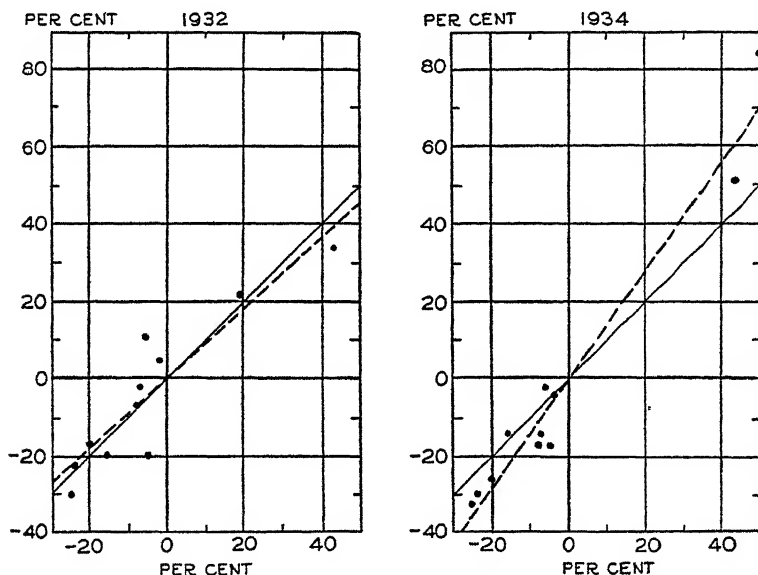
Chart 194. Seasonal Index as Percentage Deviations (on Horizontal Scale) and Percentage Deviations from 12-Month Moving Average (on Vertical Scale) together with Seasonal Amplitude Correction Line, for Receipts of Sheep and Lambs, 1932 and 1934. (For source of original data see Table 123. 1934 data are from Table 123.)

per cent. The solid line slanting at an angle of 45 degrees, with the equation $y = x$, represents the relationship that would exist between the variables if the amplitude of the seasonal index and the deviations from moving average were the same. Apparently the actual seasonal amplitude was greater than average in 1934 (since the broken line has a greater slope than the solid one); but in 1932 it was apparently slightly less than normal. These lines of relationship can be fitted by inspection (as were the Easter adjustment lines), or mathematically by the general method of fitting trend lines described on pages 395–408. The mathematical method for 1934 is shown in Table 124. Since both sets of data were made to average zero,

the estimating line passes through the point $X = 0$, $Y = 0$, and a in the usual straight line equation becomes zero.   The equation for 1932 is $y = .89x$, while that for 1934 is $y = 1.39x$.   Note that the values of b for the different years may be called *amplitude ratios*.   (Again it might parenthetically be noticed that the dots are fairly close to the line, suggesting that most of the difference between the two series is explained by a difference in amplitude.)

TABLE 124

COMPUTATION OF AMPLITUDE RATIO, AND CORRECTION OF SEASONAL INDEX FOR AMPLITUDE, FOR RECEIPTS OF SHEEP AND LAMBS, 1934

Month	Seasonal index x	Deviations from moving average y	xy	x^2	$y_c = 1.39x$	Corrected seasonal $100 + y_c$
January .	−16	−14	224	256	−22	78
February. .	−25	−32	800	625	−35	65
March	−20	−26	520	400	−28	72
April	− 7	−14	98	49	−10	90
May.. ...	− 2	− 2	4	4	− 3	97
June . .	− 8	−17	136	64	−11	89
July	− 6	− 2	12	36	− 8	92
August ..	19	19	361	361	26	126
September .	43	51	2,193	1,849	60	160
October .	50	84	4,200	2,500	70	170
November	− 5	−17	85	25	− 7	93
December	−24	−30	720	576	−33	67
Total . . .	− 1	0	9,353	6,745	− 1	1,199

Source. Table 123.

$$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{9,353}{6,745} = 1.39.$$

Equation: $y_c = 1.39x$.

5.   In order to correct our seasonal index for amplitude variations, we must substitute the different values of x in our equation; that is, *multiply each of our seasonal indexes* (taken each year as deviations from 100) by *the appropriate amplitude ratio*.   (The amplitude ratio will usually be different for each year.)   This is done in the y_c column of Table 124.   In 1934 plainly the effect of this procedure is to increase the amplitude of the seasonal index.   In 1932 the opposite result is obtained.   Finally we must *convert these y_c values into percentages averaging 100 by adding 100 to each number*, as in the last column of Table 124.

Although the procedure for 1934 only is here illustrated, each year must

be dealt with separately in similar fashion.   After the corrected seasonal of step 5 is obtained, seasonal variation is eliminated in the usual manner, by dividing the original data by the corrected seasonal.   The results of these operations can be seen in the two parts of Chart 195.   Note that
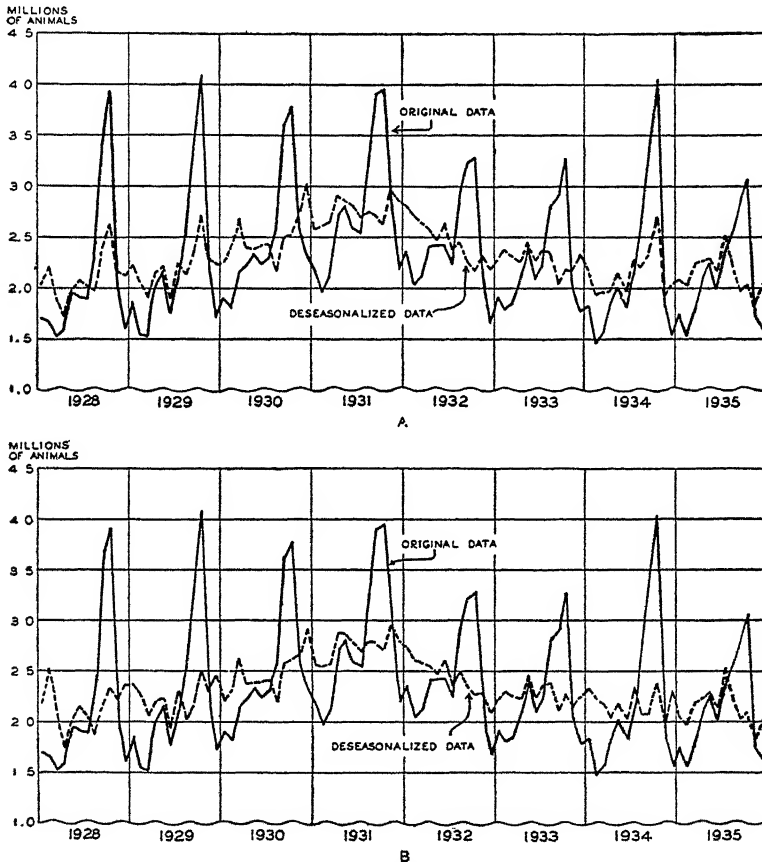
Chart 195.   Receipts of Sheep and Lambs as Deseasonalized (A) by Stable Index and (B) by Index of Varying Amplitude, 1928-1935.   (For source of original data see Table 123.)

adjustment by seasonal indexes of varying amplitude gives smoother re-sults than adjustment by a stable seasonal.

A word of caution is in order.   If a moving seasonal has been used, a change in the amplitude ratio does not necessarily indicate a change in the seasonal amplitude of the original data.   A gradual increase in the seasonal amplitude, for instance, would be reflected in the moving seasonal index rather than in the amplitude ratio; but the moving seasonal would

fail to register any sudden departures from the general trend in amplitude change.

Further Refinements of Method

Continuity of seasonal indexes. A stable seasonal index averages 100 per cent, not only for the 12-month period selected for the index, but for any consecutive 12-month period. The latter, however, is not true for any of the seasonals explained in this chapter, though in the case of a progressive or moving seasonal the discrepancy is nominal only. Particularly in the case of seasonal indexes corrected for variations in amplitude, however, the discrepancy may assume alarming proportions. The difficulty manifests itself in discontinuity of the seasonally adjusted data at the point where one year ends and the next begins. Let us assume, for instance, that the unadjusted seasonal index numbers for December 1930 and January 1931 are each 80 per cent, the amplitude adjustment to be applied, let us say, to calendar years. Now, suppose further that the amplitude ratios are .5 and 1.5 respectively. This makes the adjusted December 1930 index number 40 per cent and the January 1931 number 120. It is apparent that there will be an enormous drop in the seasonally adjusted data between December and January. Yet a little thought will convince one that the change in amplitude does not take place entirely in a month's time, but represents a transition of several months' duration.

Although there is no entirely satisfactory solution for this difficulty, one remedy, which is very laborious, is to compute an amplitude ratio for each consecutive 12-month period of the entire series. For instance, if the data ran from 1926 through 1936, the first 12-month period would run from January 1926 through December 1926, the second from February 1926 through January 1927, and so on. Altogether there would be 121 such 12-month periods and the same number of amplitude ratios. We could speak of these ratios collectively as a *moving amplitude ratio.* Following the analogy of a 12-month moving average, these ratios should be centered by a 2-month moving average, leaving 120 amplitude ratios, running from July 1926 through June 1936. The seasonal index numbers are then multiplied by these amplitude ratios to obtain the final seasonal index numbers

This procedure is laborious, but it is not entirely satisfactory. Although there is no sharp break in the continuity of the series, it has the defect that not any 12 consecutive seasonal index numbers are centered on 100 per cent. A less accurate but also much less laborious procedure than the one just described is to compute an amplitude ratio for each standard year, center the ratio on the sixth or seventh month, and interpolate arithmetically from one year to the next.

Combinations of seasonal types.   It is frequently true that the seasonal variation may be gradually changing in pattern, shifting in its timing, and varying in amplitude, or some combination of the three.   Thus, flaxseed stocks, during the crop years from July 1923 through June 1936, had 2 September peaks, 9 October peaks, and 2 November peaks, but also varied tremendously from year to year in amplitude of fluctuation, the 1927–1928 amplitude ratio being 1.69 and that for 1929–1930 being .44.   The procedure for obtaining final seasonal indexes for these data would be: (1) break data into sub-periods according to occurrence of seasonal high; (2) compute stable seasonal for each such sub-period; (3) using these seasonal indexes, compute amplitude ratios for each year (possibly using the method of interpolation described above); (4) multiply the seasonal index numbers by the appropriate amplitude ratios.

Other combinations of seasonal types require different treatment.   Considerable ingenuity is frequently required to measure and eliminate seasonal variation successfully.   Unfortunately, there is no way of telling when we have arrived at the best solution of the problem.   Complexity of procedure does not guarantee that the results obtained accurately describe the movement which we set out to measure.   Particularly if the data are originally unreliable, great refinement of method is likely to be largely wasted effort.

Correction by subtraction of seasonal.   It occasionally happens that grotesque results are obtained when seasonal is eliminated by dividing by a seasonal index.   This is especially likely to be the case when the seasonal movement typically falls almost to zero at one or more months.   Then, if in any given year the series remains materially above zero for those months, division by the extremely low seasonal index percentage will raise the deseasonalized data to a very sharp peak.   This is true to a lesser degree when the series is characterized by a single sharp seasonal peak each year.

A simple expedient is as follows.   Compute a seasonal index by whatever method seems appropriate.   The index is now converted into terms of the original data by multiplying the seasonal index numbers (expressed as percentage deviations) each year by the average value of the original series for that year.   Seasonal is then eliminated by subtracting, algebraically, the seasonal index from the original data.

It may be desirable to compute the index number, in the first instance, in such a way as to obtain a seasonal index in absolute rather than relative terms.   This will be so if the seasonal movements each year seem to be similar in absolute magnitude rather than in percentage deviations.   Inspection of a chart of the data may indicate whether this is true.   If the

evidence indicates that an index of absolute deviations should be computed, it is necessary only to adapt one of the methods with which the
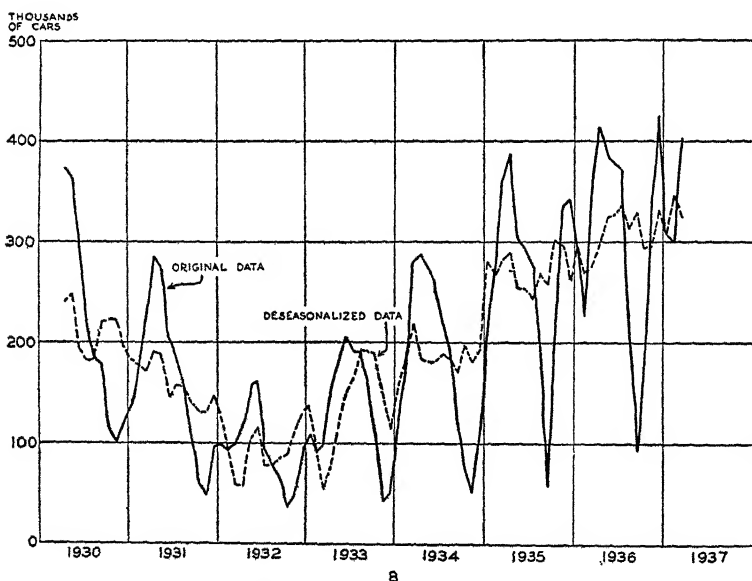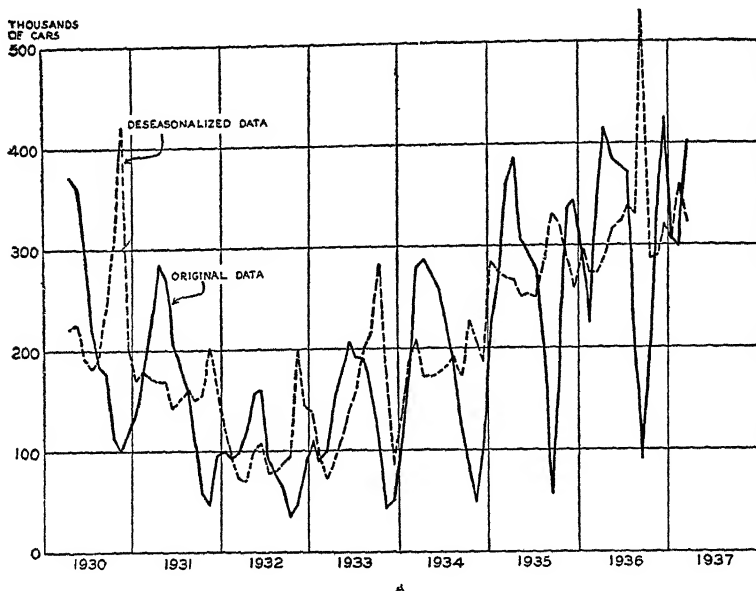
Chart 196. United States Passenger Car Production as Deseasonalized (A) by Division Method and (B) by Subtraction Method, April 1930-March 1937. (For source of original data see Chart 192.)

reader is familiar. For instance, if the moving average method is used, the moving average is subtracted from, instead of divided into, the original data; and the index from that point is constructed as usual, the final index being adjusted to total zero by the subtraction of a correction factor. Incidentally, it might be noted that any of the devices explained earlier in this chapter may be based on the subtraction method of computing seasonal. The link relative method (described in the preceding chapter) can also be adapted very easily as follows: (1) Obtain link differences by subtracting the preceding month from each month; (2) average these link differences, month by month; (3) let the first month link difference be zero, and chain the links by successive addition; (4) correct chain differences for trend by successive subtraction of correction factor; (5) adjust chain differences to total zero by subtraction of a constant correction factor.

It may be apparent, from inspection of the plotted time series, whether to compute a seasonal based on percentages or on differences, or whether to adapt the ordinary seasonal index so that the correction may be made by subtraction. However, it may be necessary to try several approaches until a satisfactory method has been found. The two parts of Chart 196 indicate the difference between the division method and the subtraction method of correcting passenger car production for seasonal. Two seasonal indexes were used in each section of the chart, the second index beginning with April 1935.

Logical basis of methods of construction. With the exception of the adjustment for Easter, the methods described in this chapter are more or less empirical in nature, depending for their validity upon the results which they produce. A method is held to be satisfactory if the deseasonalized data (1) do not show similarity of intra-year pattern (other than cyclical) in different years; (2) are not extremely irregular in their movements; and (3) are of about the same magnitude as the original data in 12-month periods.

The Easter adjustment, on the other hand, attempted to find a functional relationship between April sales minus March sales and the date of Easter. Carrying this idea further, it might be possible to find a numerical relationship over time between length of daylight and sales of incandescent lamps; or between temperature and sales of ice; or between a combination of temperature and snowfall and sales of galoshes. Computation of seasonal indexes by such a method would carry us far into the field of correlation, which is treated in the last four chapters of this book. Furthermore, it would be difficult to measure the importance, let us say, of Christmas by correlating sales with some other factor.

Intermediate between these two types of methods is that which obtains

a first approximation seasonal index by an empirical method, and then seeks to smooth this index by fitting a curve to the seasonal index numbers on the theory that the seasonal movement would present a smooth pattern if the period covered were long enough to permit an exact cancelling out of all irregular movements. Freehand smoothing of the seasonal curve is practiced by a few statisticians. The fitting of a mathematical curve is not usually advocated. Indeed, it would be easy to find logical objections to a simple curve fitted to most data. Usually there are social factors that disturb the smoothness of contour inherent in a simple mathematical curve.

Weekly Seasonal

There has been a tendency in recent years to attempt to furnish important economic data promptly and at frequent intervals, so that users can be familiar with the current situation rather than with that of a month or two ago. Naturally, weekly or daily series have rather wide irregular variations, but the calendar variations and seasonal swings also require careful attention. It is especially important to adjust such data for holidays, since one day of idleness in a week of six working days is a difference of 16.7 per cent. The steel operations data used in the following illustration, however, do not require such adjustment since they are expressed as percentages of capacity.

There is little that is new in the computation of a weekly seasonal. It is similar to the moving average method discussed in Chapter XVII. As with that method, an attempt is made to estimate values which consist of combined trend and cycle, and the original data are then expressed as percentages of these values. From these percentages the seasonal index is computed.

Trend × Cycle estimates for each week of each year are not obtained by means of a 52-week moving average (corresponding to the 12-month moving average used in computing a monthly seasonal). Since the number of days in a year is not a multiple of seven, the 52 weeks in a year end (or center) on different dates in different years. Thus the first week ending in a given year may end on January 1, 2, 3, 4, 5, 6, or 7. Because a given week may end on any one of the seven dates, there will be in general one-seventh as many observations for a week ending on a given date as there are years under consideration. Consequently, if percentages of a 52-week moving average are arrayed by dates in the year and averages taken to eliminate irregular and extraneous cyclical movements (in accordance with the method used in computing a monthly seasonal index), the observations for any week will ordinarily be too sparse to obtain a typical value for that week. Therefore a more accurate estimate of Trend ×

Cycle must be obtained—one which reaches into the cyclical peaks and troughs more faithfully. The method of making this estimate will be explained in the following paragraphs, as will the other steps involved in computing the index.[5]

Another difference between a monthly seasonal and a weekly seasonal should be noticed. Although a monthly seasonal index requires only twelve index numbers, one for each of the twelve months, a weekly seasonal index requires not merely one number for each of the 52 weeks, but 365 numbers (366 for leap years), which is one for every week ending on each possible date.

1. *Obtain monthly data.* If comparable monthly and weekly data are not available, monthly data may be obtained by taking monthly averages of the weekly data, as in column 3 of Table 125.

If the data are for weeks ending on specified dates, the results should be placed opposite the week containing the 15th of the month; if the dates specified are for the center of the week, the results should be placed opposite the date nearest the 15th.[6] The former is the procedure followed in Table 125, column 3. (In this table the years 1927 and 1936 only are shown.)

2. *Compute a monthly seasonal index.* The seasonal index is shown in column 4 of Table 125.

3. *Adjust monthly data for seasonal movement by dividing them by the seasonal index.* See starred items of column 5, Table 125.

4. *Obtain approximations of weekly values of Trend × Cycle by arithmetic interpolation of the adjusted monthly values.* See column 5. This method is not perfect, since the deseasonalized data also contain irregular movements and the interpolation process does not entirely eliminate them.

5. *Express the original weekly data as percentages of these estimates.* These percentages are supposed to contain only seasonal movement and irregular variations.

6. *Tabulate these percentages according to day of month,* as in Table 126, column 2. The items marked Δ are from column 6 of Table 125 and

---

[5] See "A Method of Calculating Weekly Seasonal Indexes," by Leroy M. Piser, *Journal of the American Statistical Association*, Vol. XXVII, September 1932, pp 307–309.

[6] Maximum accuracy is obtained for monthly averages of weekly data if the data are taken for weeks centering on specified dates, and fractional weeks at the beginning and end of each month are given fractional weights The labor of such a procedure is probably not justified by the added accuracy obtained According to the method followed in Table 125, the monthly averages do not result in estimates for calendar months, but for periods beginning and ending a few days before the end of the month. This does not impart a bias to the weekly index, however, since the weekly data are for weeks ending, rather than centering, on specified dates in the month.

TABLE 125

COMPUTATION OF PERCENTAGES OF ESTIMATED TREND × CYCLE MOVEMENTS OF RATE
OF STEEL OPERATIONS OF ENTIRE UNITED STATES INDUSTRY, 1927–1936

Year, month and day (week ending Monday.) (1)	Per cent of capacity (2)	Monthly average centered on week including 15th (3)	Monthly seasonal index (4)	Estimated Trend × Cycle (5)	Per cent of Trend × Cycle [Col 2 ÷ Col. 5] (6)
1927					
January 3	75 0				
10	76 5				
17	76.5	76 5	96 6	79 2*	96 6
24	76 5			78.7	97 2
31	78.0			78 1	99 9
February 7	79 0			77 6	101 8
14	81.0			77 0	105.2
21	83 5	82.6	108.0	76 5*	109 2
28	87 0			77 8	111 8
March 7	87.5			79 1	110.6
14	91 5			80 4	113 8
21	92 5	90 9	111.2	81 7*	113 2
28	92 0			80 6	114.1
April 4	90.0			79.5	113 2
11	88.5			78 4	112.9
18	86 5	87.3	113.0	77 3*	111.9
25	84.0			76.2	110 2
May 2	82 0			75 0	109.3
9	81 0			73 9	109 6
16	86 0	80.9	111.1	72 8*	109 9
23	81 5			72 7	112.1
30	80.0			72 6	110 2
June 7	75.5			72 4	104 3
14	74 0			72.3	102.4
21	71 0	72 9	101 0	72 2*	98.3
28	71 0			72 0	98.6
July 4	67.5			71.7	94 1
11	66.5			71 4	93 1
18	67 0	67.4	94.6	71 2	94 1
25	68.5			71 2*	96 2
August 1	68.5			71.3	96.1
8	65 5			71.4	91 7
15	66.0	66.8	93.5	71.4*	92.4
22	66.0			71.1	92 8
29	68 0			70.8	96 0

TABLE 125 (Continued)

COMPUTATION OF PERCENTAGES OF ESTIMATED TREND × CYCLE MOVEMENTS OF RATE
OF STEEL OPERATIONS OF ENTIRE UNITED STATES INDUSTRY, 1927–1936

Year, month and day (week ending Monday:)	Per cent of capacity	Monthly average centered on week including 15th	Monthly seasonal index	Estimated Trend × Cycle	Per cent of Trend × Cycle [Col 2 ÷ Col 5]
(1)	(2)	(3)	(4)	(5)	(6)
September 5	67 5			70 6	95 6
12	65 0			70.3	92.5
19	62 0	64 6	92 3	70 0*	88 6
26	64 0			69 7	91 8
October 3	65 0			69 4	93 7
10	66 0			69 0	95.7
17	64 0	65.1	94 8	68 7*	93 2
24	65 0			69 4	93 7
31	65 5			70 1	93 4
November 7	66 0			70.8	93 2
14	67 0			71 5	93.7
21	68 5	66.9	92.6	72 2*	94 9
28	66 0			72 1	91 5
December 5	61 0			72 0	84.7
12	63 5			71 8	88 4
19	67 5	65 5	91 4	71 7*	94.1
26	70.0			73.1	95 8
1936 January 6	48 0			55.1	87.1
13	51 0			53 6	95 1
20	51 0	50.3	96.6	52.1*	97.9
27	51 0			51.2	99.6
February 3	50 5			50 3	100.4
10	52.0			49 4	105.3
17	53.0	52.4	108.0	48.5*	109 3
24	54.0			48.9	110.4
March 2	55.0			49.3	111.6
9	56.0			49.7	112.7
16	58 0	55.7	111.2	50.1*	115 8
23	50 5			52.0	97.1
30	59.0			53 9	109.5
April 6	63.0			55 8	112.9
13	66 0			57 7	114.4
20	70.0	67.4	113.0	59 6*	117.4
27	70.5			60.2	117.1

TABLE 125 (Continued)

COMPUTATION OF PERCENTAGES OF ESTIMATED TREND × CYCLE MOVEMENTS OF RATE OF STEEL OPERATIONS OF ENTIRE UNITED STATES INDUSTRY, 1927–1936

Year, month and day (week ending Monday.) (1)	Per cent of capacity (2)	Monthly average centered on week including 15th (3)	Monthly seasonal index (4)	Estimated Trend × Cycle (5)	Per cent of Trend × Cycle [Col 2 ÷ Col 5] (6)
May 4	70 0			60 9	114 9
11	69 0			61 6	112 0
18	69 0	69 1	111 1	62.2*	110 9
25	68 5			64 0	107.0
June 1	68 5			65 9	103.9
8	69 5			67 7	103.7
15	70 5	70 2	101 0	69 5*	101 4
22	71 5			70 1	102 0
29	71 5			70 7	101 1
July 6	65 5			71 3	91 9
13	67 0			71.9	93 2
20	70 0	68 6	94 6	72 5*	96 6
27	72 0			73 6	97 8
August 3	72 0			74.6	96.5
10	71 5			75.7	94.5
17	70 5	71.8	93.5	76.8*	91.8
24	72.5			77 0	94 2
31	72.5			77.3	93.8
September 7	69 0			77 5	89 0
14	71 0			77.8	91 3
21	73 5	72 0	92 3	78 0*	94 2
28	74.5			78.3	95 1
October 5	75.5			78.6	96 1
12	75.5			78.8	95 8
19	75 0	75.0	94.8	79.1*	94 8
26	74 0			79 3	93 3
November 2	74 0			79.6	93 0
9	74 5			79 8	93.4
16	74 5	74 5	92 6	80 5*	92.5
23	74.5			81 1	91 9
30	75.0			81 7	91.8
December 7	77.0			82 2	93.7
14	80.0			82.8	96.6
21	81.0	76 2	91.4	83.4*	97.1
28	68 0				

* These items are Col 3 ÷ Col 4.   Other items in column are obtained by interpolation
Source: Revised data furnished to writers by Standard Statistics Co.

TABLE 126

COMPUTATION OF WEEKLY SEASONAL OF RATE OF STEEL OPERATIONS, 1927–1936

Month and day (1)	Percentages of Trend × Cycle (2)		Averages and inter-polations (3)	Moving modified mean† (4)	Smoothed moving mean (5)	Weekly seasonal [Col. 5 × 1.00221]‡ (6)
January						
1	89.3		89.3	87 6	87.5	87.7
2	90.1	81 3	85.7	86 6	86 5	86 7
3	..		84 8	85.9	86 0	86 2
4	84 0		84.0	85.6	85 5	85 7
5	87 2		87.2	86 1	86 0	86.2
6	85 6	87 1Δ	86.4	88.7	88 5	88 7*
7	95.5	93 3	94.4	90 9	91 0	91 2
8	92 5		92 5	92 9	92 0	92 3
9	93.7	92.3	93.0	93.2	92 5	92.7
10	..		93.2	93.2	93 0	93 2
11	93 5		93.5	93 7	93.5	93 7
12	94 8		94 8	94.2	94 0	94 2
13	93 8	95 1Δ	94.4	95 1	95.0	95.2*
14	94 4	98 1	96 2	95 9	96 0	96 2
15	97.3		97 3	96 5	96.5	96 7
16	95 9	97 6	96 8	96.9	97 0	97 2
17	96.6Δ		96 6	97 8	98 0	98 2
18	99.2		99.2	98 3	98.5	98 7
19	103 2		103.2	98 9	99 0	99 2
20	99 7	97 9Δ	98 8	99.0	99.5	99 7*
21	96.3	101.4	98 8	99.6	99 5	99 7
22	99 1		99.1	98 9	99.5	99 7
23	99.6	102.3	101 0	99.6	100.0	100.2
24	97 2Δ		97 2	101 7	101.5	101 7
25	110 9		110 9	102.7	102 5	102 7
26	105 0		105.0	103.6	103.0	103.2
27	104 7	99 6Δ	102 2	103 6	103.0	103 2*
28	99 5	107.9	103.7	103.6	103 0	103 2
29	96.6		96 6	101.9	103 0	103.2
30	108.4	106.9	107 6	102 9	103.0	103 2
31	99.9Δ		99.9	103.5	103.5	103.7
February						
1	105.2		105.2	104.8	104.0	104 2
2	105.4		105.4	104 8	104 5	104.7
3	107.3	100.4Δ	103 8	104 8	105 0	105 2*
4	102.4	110 2	106.3	105 2	105.0	105.2
5	101.4		101.4	104.0	105.5	105.7
6	108.2	108.6	108.4	105.5	106.0	106 2
7	101.8Δ		101 8	106.4	106.5	106.7
8	108 9		108 9	108.3	107.5	107.7
9	109.0		109.0	108.2	108.0	108 2

533

TABLE 126 (Continued)

COMPUTATION OF WEEKLY SEASONAL OF RATE OF STEEL OPERATIONS, 1927–1936

Month and day (1)	Percentages of Trend × Cycle (2)		Averages and interpolations (3)	Moving modified mean† (4)	Smoothed moving mean (5)	Weekly seasonal [Col 5 × 1 00221]# (6)
February						
10	109 8	105 3Δ	107 6	108 2	108 0	108 2*
11	105 3	110.7	108 0	108.2	108 0	108 2
12	107 0		107 0	107 5	108 0	108 2
13	109 3	110.2	109 8	108.0	108 0	108 2
14	105.2Δ		105 2	108 5	108 5	108 7
15	109.1		109.1	109 5	109 0	109 2
16	109 5		109 5	108 9	109 5	109 7
17	110 4	109 3Δ	109 8	109 5	109 5	109 7*
18	108 9	107 0	108.0	109 7	109 5	109 7
19	111 1		111.1	109 6	109.5	109 7
20	107 8	111.7	109.8	109 0	109 5	109 7
21	109 2Δ		109 2	109.9	109 5	109 7
22	104 2		104.2	109 9	109 5	109.7
23	110 6		110 6	109 1	109 5	109.7
24	111 1	110 4Δ	110 8	109.6	109.5	109 7*
25	109 0	105 9	107.4	110 2	110 0	110 2
26	115 4		115.4	110.6	110 5	110 7
27	108 2	110 1	109 2	110 2	110 5	110 7
28	111 8Δ		111 8	110.6	110 5	110 7
December						
1	91 1		91.1	91 6	91 5	91.7
2	90 3	96.7	93 5	91 3	91.5	91 7
3	94.0	86.1	90 0	90.3	90 5	90.7
4	89 7		89.7	90 4	90 5	90.7
5	84 7Δ	95.0	89 8	90 4	90 5	90.7
6			91 5	90.7	90 5	90.7
7	92 7	93 7Δ	93.2	91 7	91 0	91.2*
8	90 9		90.9	91 7	91 5	91 7
9	88.9	96 6	92 8	92.0	91.5	91.7
10	97.2	88 7	90.0	91.1	91 5	91 7
11	92 3		92 3	91.3	91 5	91 7
12	88.4Δ	90.6	89.5	91 3	92 0	92 2
13			91 6	92 6	92 5	92.7
14	90.9	96.6Δ	93.8	92 6	92 5	92.7*
15	95.6		95.6	92 6	92 5	92.7
16	90.8	94.0	92 4	93 9	93 0	93 2
17	88.9	91.2	90.0	93 2	93 0	93 2
18	97 3		97.3	92 3	93 0	93.2
19	94.1Δ	89.0	91.6	92.9	93.0	93.2

TABLE 126 (Continued)

COMPUTATION OF WEEKLY SEASONAL OF RATE OF STEEL OPERATIONS, 1927–1936

Month and day (1)	Percentages of Trend × Cycle (2)		Averages and inter- polations (3)	Moving modified mean† (4)	Smoothed moving mean (5)	Weekly seasonal [Col. 5 × 1 00221]# (6)
20			92.9	92.9	93 0	93.2
21	91.3	97 1△	94.2	92 0	92 5	92 7*
22	85 9		85 9	92 2	92 0	92 2
23	90 3	92 9	91 6	92.6	91.5	91.7
24	90.7	93 2	92 0	89.9	91 0	91.2
25	97.6		97 6	89 9	90 0	90 2
26	95 8△	76 2	86.0	86 6	86 5	86.7
27			81 9	81 9	82 0	82.2
28	77 8		77 8	81 9	82 0	82.2*
29	74 1		74 1	82 5	82 5	82 7
30	89 1	86 6	87 8	85 0	85 0	85 2
31	94 7	92 6	93 6	87 6	87.5	87 7
Total	..			.	36,419 5	36,493 8‡

△ From column 6, Table 125.
* These index numbers are for w~~k~ ~~~ ~~~~ in 1936　See also Table 127
† Mean of middle three of five ·　·　（　　'
# Correction factor: 1 00221 = 36,500 0 − 36,419 5
‡ Discrepancy between 36,500 0 and 36,493 8 is due to rounding. It represents about .02 of one per cent.
Source: See Table 125.

refer to the years 1927 and 1936. Table 125 would include the other values, also, if all intervening years were shown. (In Table 126, January, February, and December only are shown.) It should be noticed that in some instances there will be more than one value for a particular day; in other instances there will be none.

7. *Obtain one value for each day of each month.* This is done in column 3 of Table 126 by averaging when there are two or more values, and inter- polating when there are none.

8. *Smooth these values by a moving average.* In column 4 of Table 126 a modified mean—the middle three of five arrayed items—was used. Using a small number of items makes for flexibility, while the modified mean makes for smoothness by toning down the influence of extremes, which are likely to be great with weekly data. The type of moving average to use, however, must be decided separately for each series. Chart 197 shows by dots the percentages of column 2 of Table 126. The thin broken line represents the moving modified mean.

9. *Further smooth the values by inspection.* This is shown by the heavy solid line of Chart 197. The logic of this double smoothing process is that, since there are only one or two items for any day, irregularities cannot
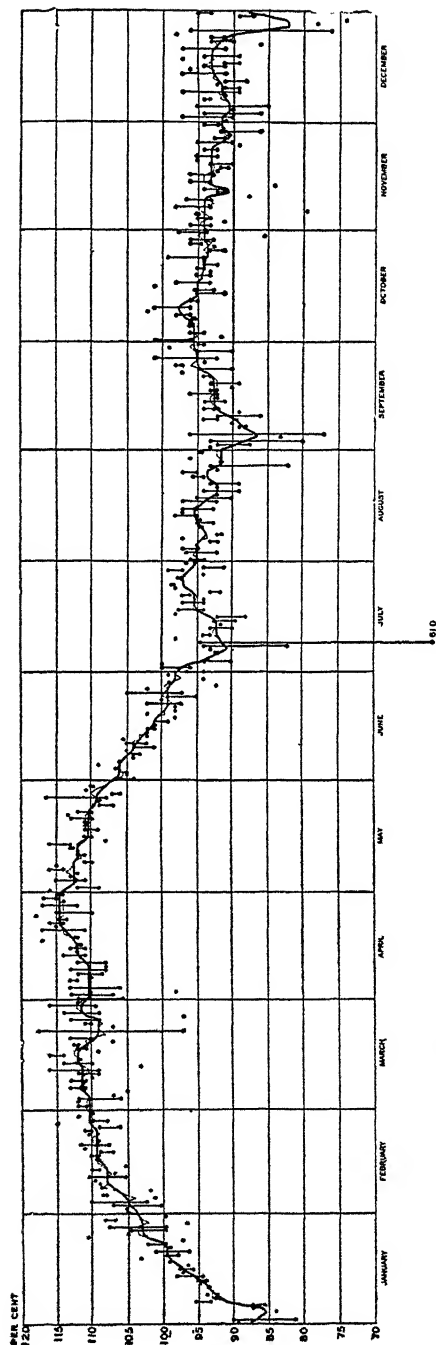
Chart 197. Derivation of Weekly Seasonal Index of Rate of Steel Operations in the United States, 1927-1936. (Dots represent percentages of estimated $T \times C$. Thin vertical lines connect two or more observations occurring on the same day of the year. The thin broken line is the moving modified mean. The heavy solid line is the smoothed moving modified mean. Data of Table 126.)

TABLE 127

SEASONAL ADJUSTMENT OF WEEKLY MOVEMENTS OF RATE OF STEEL OPERATIONS IN THE UNITED STATES, 1936

Week ended Monday: (1)	Per cent of capacity (2)	Weekly seasonal* (3)	Deseasonalized data [Col 2 ÷ Col 3] (4)	Week ended Monday: (1)	Per cent of capacity (2)	Weekly seasonal* (3)	Deseasonalized data [Col 2 ÷ Col 3] (4)
January				July			
6	48.0	88 7	54 1	6	65.5	91.7	71.4
13	51 0	95 2	53 6	13	67 0	92.2	72.7
20	51.0	99.7	51.2	20	70 0	95.7	73.1
27	51 0	103.2	49.4	27	72.0	97.2	74 1
February				August			
3	50.5	105.2	48 0	3	72 0	95.2	75.6
10	52.0	108.2	48 1	10	71.5	94.2	75 9
17	53.0	109.7	48.3	17	70 5	94.2	74 8
24	54.0	109 7	49.2	24	72 5	93.7	77.4
				31	72.5	91.7	79.1
March				September			
2	55 0	110.7	49.7	7	69 0	88.7	·77.8
9	56.0	111.7	50.1	14	71 0	92.7	76 6
16	58.0	112.2	51.7	21	73 5	93.2	78.9
23	50 5	109.2	46.2	28	74.5	95.7	77.8
30	59.0	111.7	52.8				
April				October			
6	63.0	110.7	56.9	5	75.5	95.7	78 9
13	66.0	112.2	58 8	12	75.5	96.2	78.5
20	70 0	114.3	61.2	19	75.0	94 7	79.2
27	70.5	114.8	61.4	26	74.0	93.7	79.0
May				November			
4	70.0	112.7	62.1	2	74.0	94.2	78.6
11	69 0	112.2	61.5	9	74.5	94.2	79.1
18	69 0	110.7	62.3	16	74.5	93 2	79.9
25	68.5	109.7	62.4	23	74.5	93.2	79.9
				30	75.0	91.7	81.8
June				December			
1	68.5	106.7	64.2	7	77 0	91.2	84.4
8	69.5	104.7	66.4	14	80.0	92.7	86.3
15	70.5	101.7	69.3	21	81.0	92.7	87.4
22	71.5	99.2	72.1	28	68.0	82.2	82.7
29	71.5	98 2	72.8				

* These are the starred items of Table 126 column 6.
Source: Tables 125 and 126.

completely cancel out, even by the use of a moving modified mean, such as is used in step 8.

10. *Adjust the values to total 36,500.* This must be done because the seasonal index should average 100, and there are 365 days in a year (except leap year). The method is to divide 36,500 by the total of the values of column 5 (Table 126) and multiply each value by the quotient. We now have a seasonal index number for a week ending on every day of the year. Chart 197 does not show a separate line for the final weekly seasonal, since it would be indistinguishable from the solid line.

Chart 142 affords a comparison between the monthly and weekly sea-
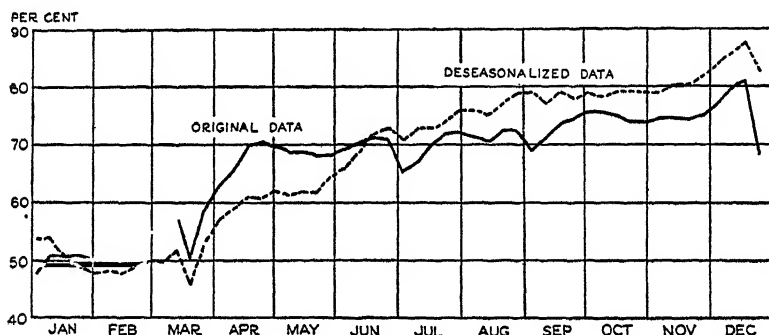
Chart 198. Rate of Steel Operations in the United States Before and After Adjustment for Seasonal Variation, by Weeks, 1936. (Data of Table 127)

sonal indexes of these data. It is apparent that the monthly index is not sufficiently flexible to be used with weekly data.

The deseasonalizing of weekly data presents no new problem. The process is illustrated by Table 127, for the year 1936. The seasonal index numbers of column 3 (Table 127) are taken from column 6 of Table 126. The original data are now divided by these numbers and the results recorded in Table 127, column 4. They are shown graphically in Chart 198. Note that the larger irregularities are reduced somewhat, and the improvement in steel operations during the year is more marked than is apparent from the original data.

Selected References

G. R. Davies and Dale Yoder: *Business Statistics*, pages 241-249; John Wiley and Sons, New York, 1937.
Simon Kuznets: *Seasonal Variations in Industry and Trade;* National Bureau of Economic Research, New York, 1933.

Horst Menderhausen. "Annual Survey of Statistical Technique Methods of Computing and Eliminating Changing Seasonal Fluctuations," *Econometrica*, Volume V, 1937, pages 234–262.

F. C. Mills. *Statistical Methods Applied to Economics and Business* (Revised Edition), pages 522–529; Holry Holt and Co , New York, 1938. The use of analysis of variance for testing significance of seasonal fluctuations is illustrated.

CHAPTER XIX

CYCLICAL MOVEMENTS

Five chapters on time series analysis have been presented for the reader's study. Chapter XIV explained that economic time series were typically the product of secular trend (T), cyclical movements (C), seasonal variations (S), and irregular fluctuations (I). Chapters XV and XVI were devoted to a consideration of types of trends, how to select the appropriate type, methods of trend fitting, and how to remove statistically from the data the effect of trend. Chapters XVII and XVIII, similarly, considered types of seasonal variations, their measurement and elimination. In this chapter we shall consider methods of measuring the time series movement that is probably of most interest to economists: cyclical variation.

There are several methods which are sometimes used, and we shall consider them in the following order: (1) residual method; (2) direct method; (3) harmonic analysis; (4) method of cyclical averages.

Residual Method

This is the orthodox method, and the one most commonly used. It consists in successively eliminating seasonal and trend from the data, thus obtaining cyclical-irregular movements, and then perhaps further smoothing the results to obtain the cyclical movements, or *cyclical relatives*, as they are sometimes designated. It will be recalled that in Chapter XVII, Table 116 illustrated the elimination of seasonal by division of original data by the seasonal index. Consequently this process need not be repeated here. Since the seasonal index numbers are percentages averaging 100 per cent each year, the deseasonalized data are in terms of original units and of approximately the same magnitude. Cyclical-irregular movements are now obtained by dividing the seasonally adjusted data month by month by the trend values, obtaining percentages, as in column 4 of Table 128. Trend values are obtained by the process described on pages 395–411, using the trend equation shown at the top of page 409. The cyclical movements of column 5 are the result of smoothing the data of column 4 by a moving average, as will shortly be explained.

TABLE 128
COMPUTATION OF CYCLICAL MOVEMENTS FROM DESEASONALIZED UNITED STATES
MAGAZINE ADVERTISING DATA, 1921–1937

Year and month	Deseasonalized data	Trend values	Cyclical-irregular movements (per cent) [Col 2 ÷ Col 3]	Cyclical relative (per cent) 3-month binomial moving average of column 4
(1)	(2)	(3)	(4)	(5)
1921				
January......... .	2,334	2,166	107 8	
February........	2,038	2,176	93 7	95 3
March...	1,881	2,186	86 0	86 6
April	1,776	2,196	80.9	83.4
May.......... .	1,890	2,206	85 7	84 3
June.............	1,884	2,216	85 0	85 6
July..............	1,928	2,226	86 6	86 3
August.........	1,947	2,236	87 1	85 0
September	1,776	2,246	79 1	79 4
October....	1,630	2,255	72 3	74 3
November	1,668	2,265	73 6	73 6
December	1,700	2,275	74 7	76.8
1922				
January..	1,925	2,285	84.2	80 6
February... . .	1,819	2,295	79 3	80 2
March....	1,803	2,305	78 2	78 8
April	1,837	2,315	79 4	80 2
May........	1,952	2,325	84 0	83 2
June...... ..	1,994	2,335	85 4	86 0
July	2,089	2,345	89 1	89 0
August.. . .	2,175	2,354	92 4	91.0
September... ..	2,127	2,364	90 0	91 4
October.	2,207	2,374	93 0	91.6
November	2,161	2,384	90 6	92.8
December..........	2,318	2,394	96.8	96.7
1923				
January...........	2,468	2,404	102 7	100.1
February.........	2,367	2,414	98 1	99.5
March.............	2,399	2,424	99 0	99.8
April	2,507	2,434	103 0	101.9
May..............	2,511	2,444	102 7	103.2
June.............	2,565	2,454	104.5	104 7
July............. .	2,635	2,463	107 0	105 8
August............	2,589	2,473	104.7	104.2
September.........	2,497	2,483	100 6	102 2
October...........	2,567	2,493	103.0	102.0
November.........	2,542	2,503	101.6	102.5
December..........	2,609	2,513	103.8	104.0
1924				
January...........	2,690	2,523	106.6	105.2
February..........	2,631	2,533	103.9	105.5
March.............	2,735	2,543	107 6	106 9
April.............	2,769	2,553	108.5	106.8
May..............	2,633	2,562	102.8	105.3
June.............	2,758	2,572	107.2	103.2
July.............	2,472	2,582	95.7	99.0
August............	2,526	2,592	97.5	96.3
September........	2,458	2,602	94.5	95.4
October..........	2,483	2,612	95.1	95.4
November.........	2,542	2,622	96 9	97.6
December.........	2,673	2,632	101.6	98.6

TABLE 128 (Continued)
COMPUTATION OF CYCLICAL MOVEMENTS FROM DESEASONALIZED UNITED STATES
MAGAZINE ADVERTISING DATA, 1921–1937

Year and month	Deseasonalized data	Trend values	Cyclical-irregular movements (per cent) [Col 2 ÷ Col 3]	Cyclical relatives (per cent) 3-month binomial moving average of column 4
(1)	(2)	(3)	(4)	(5)
1925				
January	2,489	2,642	94.2	96 9
February	2,585	2,652	97 5	96 2
March	2,544	2,662	95 6	95 6
April	2,497	2,671	93 5	94 1
May	2,515	2,681	93 8	94 1
June	2,568	2,691	95 4	94 6
July	2,534	2,701	93.8	94 8
August	2,608	2,711	96 2	96 6
September	2,725	2,721	100 1	99 2
October	2,736	2,731	100 2	101 0
November	2,843	2,741	103.7	102 2
December	2,785	2,751	101 2	102 0
1926				
January	2,812	2,761	101 8	102 7
February	2,936	2,770	106 0	104 0
March	2,842	2,780	102 2	103 0
April	2,828	2,790	101 4	101 7
May	2,851	2,800	101 8	102 5
June	2,947	2,810	104 9	103 6
July	2,902	2,820	102 9	104 4
August	3,029	2,830	107 0	106 6
September	3,104	2,840	109 3	108 4
October	3,078	2,850	108 0	108 8
November	3,143	2,860	109 9	108 2
December	3,019	2,870	105.2	106.0
1927				
January	2,991	2,879	103.9	105 0
February	3,092	2,889	107 0	105 8
March	3,053	2,899	105.3	104 8
April	2,959	2,909	101.7	104.2
May	3,152	2,919	108 0	104.5
June	2,936	2,929	100 2	102.3
July	2,966	2,939	100.9	101 8
August	3,096	2,949	105 0	103 3
September	3,029	2,959	102 4	103.1
October	3,048	2,969	102.7	103.2
November	3,131	2,978	105.1	103.2
December	2,988	2,988	100.0	100.4
1928				
January	2,886	2,998	96.3	97.5
February	2,928	3,008	97.3	97 7
March	3,013	3,018	99.8	99 9
April	3,109	3,028	102.7	101 2
May	3,026	3,038	99.6	100 0
June	2,983	3,048	97.9	99.7
July	3,165	3,058	103.5	100.6
August	2,997	3,068	97.7	99 7
September	3,076	3,078	99.9	99 7
October	3,127	3,087	101.3	100 3
November	3,055	3,097	98.6	98.6
December	2,984	3,107	96.0	98.0

TABLE 128 (Continued)
COMPUTATION OF CYCLICAL MOVEMENTS FROM DESEASONALIZED UNITED STATES
MAGAZINE ADVERTISING DATA, 1921–1937

Year and month	Deseasonalized data	Trend values	Cyclical-irregular movements (per cent) [Col 2 ÷ Col 3]	Cyclical relatives (per cent) 3-month binomial moving average of column 4
(1)	(2)	(3)	(4)	(5)
1929				
January..	3,165	3,117	101.5	100.7
February.	3,249	3,127	103.9	104.2
March	3,378	3,137	107 7	107.2
April ...	3,453	3,147	109 7	108 8
May.	3,414	3,157	108 1	108 8
June	3,457	3,167	109.2	109.2
July . .	3,510	3,177	110 5	109 0
August	3,375	3,186	105 9	107.7
September.. ...	3,467	3,196	108 5	106.9
October.	3,360	3,206	104.8	105.6
November .	3,355	3,216	104.3	104.7
December	3,401	3,226	105.4	104.6
1930				
January. ..	3,336	3,236	103 1	102.1
February . .	3,143	3,246	96 8	98.6
March . .	3,178	3,256	97 6	97.0
April.	3,137	3,266	96.1	95.2
May.. . ..	2,983	3,276	91 1	92.6
June	3,019	3,286	91 9	91.0
July	2,942	3,295	89 3	89.3
August.	2,869	3,305	86.8	88.1
September. .	2,962	3,315	89.4	87.4
October.	2,800	3,325	84 2	85.1
November... .	2,753	3,335	82.5	82.7
December	2,730	3,345	81.6	81.3
1931				
January....... ..	2,664	3,355	79 4	79.7
February..	2,639	3,365	78.4	78.1
March..	2,569	3,375	76 1	75.7
April	2,448	3,385	72 3	73.1
May...	2,435	3,394	71.7	72.0
June.......... ..	2,459	3,404	72.2	71.6
July.	2,399	3,414	70 3	70.6
August.	2,389	3,424	69 8	69.6
September	2,359	3,434	68 7	68.5
October....	2,298	3,444	66 7	66.5
November..... ..	2,212	3,454	64 0	63.8
December...... ..	2,101	3,464	60.7	61.4
1932				
January..... ...	2,091	3,474	60.2	60.2
February..........	2,079	3,484	59.7	59.4
March	2,032	3,494	58.2	57.6
April	1,900	3,503	54.2	54.9
May	1,867	3,513	53.1	52.2
June.............	1,713	3,523	48 6	49.4
July.............	1,673	3,533	47 4	47.4
August..	1,636	3,543	46 2	45.4
September........	1,494	3,553	42 0	43.0
October	1,489	3,563	41.8	42.5
November........	1,587	3,573	44 4	43.7
December	1,589	3,583	44 3	43.6

TABLE 128 (Continued)
COMPUTATION OF CYCLICAL MOVEMENTS FROM DESEASONALIZED UNITED STATES
MAGAZINE ADVERTISING DATA, 1921–1937

Year and month	Deseasonalized data	Trend values	Cyclical-irregular movements (per cent) [Col 2 ÷ Col 3]	Cyclical relatives (per cent) 3-month binomial moving average of column 4
(1)	(2)	(3)	(4)	(5)
1933				
January	1,486	3,593	41.4	42 5
February	1,549	3,602	43 0	42.4
March	1,516	3,612	42 0	41 4
April	1,399	3,622	38 6	39.6
May	1,420	3,632	39 1	38.8
June	1,390	3,642	38 2	39.3
July	1,527	3,652	41.8	41.7
August	1,651	3,662	45.1	43.9
September	1,604	3,672	43.7	44 9
October	1,733	3,682	47 1	46.1
November	1,719	3,692	46.6	46.8
December	1,734	3,702	46.8	47.4
1934				
January	1,831	3,711	49.3	48 7
February	1,835	3,721	49 3	49.5
March	1,873	3,731	50.2	50 8
April	1,998	3,741	53.4	52 9
May	2,050	3,751	54.7	54.3
June	2,044	3,761	54 3	55.6
July	2,224	3,771	59.0	57.2
August	2,139	3,781	56.6	56.8
September	2,083	3,791	54.9	55.4
October	2,098	3,801	55.2	55.1
November	2,097	3,810	55.0	54.8
December	2,068	3,820	54.1	54.6
1935				
January	2,105	3,830	55.0	54.6
February	2,094	3,840	54.5	54.8
March	2,117	3,850	55.0	55.3
April	2,184	3,860	56.6	55.9
May	2,146	3,870	55 5	55.4
June	2,102	3,880	54.2	55 1
July	2,198	3,890	56 5	55 2
August	2,088	3,900	53.5	54.1
September	2,066	3,910	52 8	52.7
October	2,021	3,919	51.6	51.7
November	1,992	3,929	50.7	52.6
December	2,259	3,939	57.3	55.6
1936				
January	2,258	3,949	57.2	56.9
February	2,212	3,959	55.9	57.0
March	2,336	3,969	58.9	58.0
April	2,314	3,979	58 2	58.5
May	2,338	3,989	58.6	58 7
June	2,374	3,999	59.4	59.1
July	2,361	4,009	58.9	59.0
August	2,364	4,018	58.8	58.9
September	2,376	4,028	59.0	59.3
October	2,444	4,038	60.5	60 3
November	2,476	4,048	61.2	62.0
December	2,644	4,058	65.2	64.5

TABLE 128 (Continued)
COMPUTATION OF CYCLICAL MOVEMENTS FROM DESEASONALIZED UNITED STATES
MAGAZINE ADVERTISING DATA, 1921–1937

Year and month	Deseasonalized data	Trend values	Cyclical-irregular movements (per cent) [Col 2 − Col 3]	Cyclical relatives (per cent) 3-month binomial moving average of column 4
(1)	(2)	(3)	(4)	(5)
1937				
January........ ..	2,704	4,068	66.5	64 8
February.	2,494	4,078	61 2	62 9
March	2,569	4,088	62.8	62.5
April . .	2,594	4,098	63.3	63 6
May	2,670	4,108	65 0	64.8
June	2,721	4,118	66.1	65 6
July	2,683	4,127	65 0	66 0
August	2,815	4,137	68 0	66 6
September... .	2,717	4,147	65 5	65 6
October	2,643	4,157	63 6	64 4
November. . .	2,705	4,167	64 9	65.1
December ...	2,801	4,176	67 1	

Source: Deseasonalized data are from Table 116, trend equation is from Table 89

Remembering our assumption that original data $= T \times C \times S \times I$, we may describe our process algebraically as follows:

Deseasonalized data: $$\frac{T \times C \times S \times I}{S} = T \times C \times I.$$

Cyclical-irregular movements: $$\frac{T \times C \times I}{T} = C \times I.$$

The data before and after removing trend are shown by the solid lines of Charts 199 parts A and B.

It makes no difference whether seasonal is eliminated first, and then trend, or whether the order of elimination is reversed. Thus we may write:

Data adjusted for trend: $$\frac{T \times C \times S \times I}{T} = C \times S \times I.$$

Cyclical-irregular movements: $$\frac{C \times S \times I}{S} = C \times I.$$

Either of these variations of procedure might be called a method of successive elimination, since T and S are successively eliminated, leaving $C \times I$ as a residual.

Still a third variation of the residual method is possible. The term "normal" is frequently and variously used in economics and statistics. Thus, from a long-run point of view it is normal for industry to increase steadily, and from a short-run viewpoint it is normal for business to vary with the season of the year. A more comprehensive view is that both

movements together are "normal" Thus, defining normal as $T \times S$, we may obtain percentages of normal by dividing the original data by $T \times S$, and so obtain cyclical-irregular movements.

Cyclical-irregular movements: $\dfrac{T \times C \times S \times I}{T \times S} = C \times I.$

The three variations of the residual method are illustrated for the year
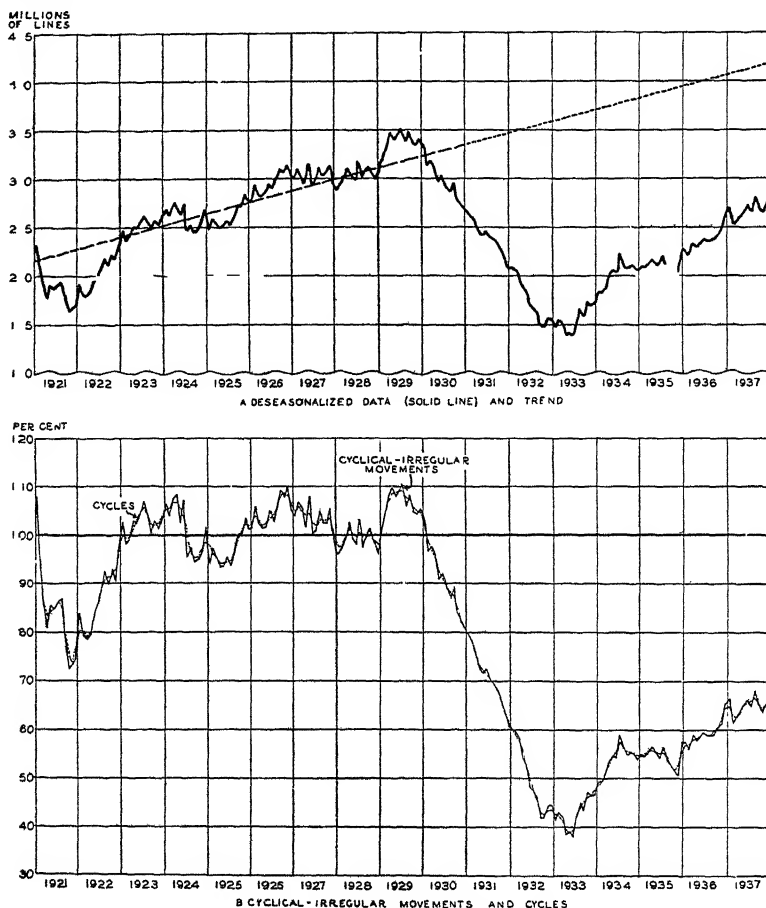
Chart 199. Deseasonalized Data, Trend, Cyclical-Irregular Movements, and Cycles, United States Magazine Advertising, 1921-1937. (Data of Table 128.)

1936 in the three sections of Table 129. Note that, except for an occasional minor discrepancy in the last digit (due to rounding), the final results are identical for each procedure. Which of the three methods to use is a matter of convenience. Probably the first method is the most

TABLE 129

THREE ALTERNATIVE METHODS OF DERIVING CYCLICAL-IRREGULAR MOVEMENTS IN UNITED STATES MAGAZINE ADVERTISING, 1936

(Original data in thousands of lines)

Method A

Year and month	Original data $T \times C \times S \times I$	Seasonal index (per cent) S	Deseasonalized data $T \times C \times I$ $[(T \times C \times S \times I) \div S]$	Trend values T	Cyclical-irregular percentages $C \times I$ $[(T \times C \times I) \div T]$
January ..	1,696	75 1	2,258	3,949	57 2
February .	2,128	96 2	2,212	3,959	55 9
March .	2,511	107 5	2,336	3,969	58 9
April.	2,860	123 6	2,314	3,979	58 2
May. .	2,852	122 0	2,338	3,989	58 6
June	2,637	111 1	2,374	3,999	59 4
July ..	1,967	83 3	2,361	4,009	58 9
August	1,695	71 7	2,364	4,018	58 8
September	2,084	87.7	2,376	4,028	59 0
October	2,637	107 9	2,444	4,038	60 5
November .	2,736	110 5	2,476	4,048	61.2
December	2,731	103 3	2,644	4,058	65 2

Method B

Year and month	Original data $T \times C \times S \times I$	Trend values T	Per cent of trend $C \times S \times I$ $[(T \times C \times S \times I) \div T]$	Seasonal index (per cent) S	Cyclical-irregular percentages $C \times I$ $[(C \times S \times I) \div S]$
January	1,696	3,949	42 9	75 1	57 1
February .	2,128	3,959	53 8	96 2	55 9
March .	2,511	3,969	63.3	107 5	58.9
April	2,860	3,979	71.9	123 6	58 2
May	2,852	3,989	71.5	122.0	58.6
June . .	2,637	3,999	65 9	111 1	59 3
July	1,967	4,009	49 1	83 3	58 9
August	1,695	4,018	42 2	71.7	58.9
September .	2,084	4,028	51.7	87.7	59 0
October. ..	2,637	4,038	65.3	107 9	60.5
November .	2,736	4,048	67 6	110 5	61.2
December ..	2,731	4,058	67.3	103 3	65 2

Method C

Year and month	Original data $T \times C \times S \times I$	Trend values T	Seasonal index (per cent) S	"Normal" values $T \times S$	Cyclical-irregular percentages $C \times I$ $[(T \times C \times S \times I) \div (T \times S)]$
January.	1,696	3,949	75.1	2,966	57.2
February.....	2,128	3,959	96.2	3,809	55.9
March.......	2,511	3,969	107.5	4,267	58.8
April........	2,860	3,979	123 6	4,918	58.2
May.........	2,852	3,989	122.0	4,867	58.6
June........	2,637	3,999	111.1	4,443	59.4
July........	1,967	4,009	83 3	3,339	58.9
August	1,695	4,018	71.7	2,881	58.8
September...	2,084	4,028	87.7	3,533	59.0
October......	2,637	4,038	107.9	4,357	60.5
November....	2,736	4,048	110 5	4,473	61.2
December. .	2,731	4,058	103.3	4,192	65.1

Source: Original data are from Table 108 Other values are computed from those data

common, since it may frequently be desired to study separately the seasonally adjusted data, but only rarely data adjusted for trend alone.   On the other hand, if a seasonal index is to be obtained by averaging percentages of trend, it is convenient to eliminate trend at the outset.   However, if the sole object of the analysis is to obtain cyclical relatives, it will be easiest to utilize the third method, which substitutes a multiplication for a division.   Chart 199B shows by the solid line the cyclical-irregular movements which result from any of these procedures.

Reducing minor irregularities. Although the curve of Chart 199B is remarkably regular, there are still a few minor irregularities which appear. These are in the main attributable to the interplay of a multitude of forces other than those being analyzed.   To a slight degree they may be due to the fact that our seasonal index is not perfect.   There is no entirely satisfactory method of eliminating these fluctuations.   However, by the use of a moving average the curve can be smoothed so as to bring the cyclical movements into clearer relief.   If the analyst is interested primarily in the combined trend and cycle, the seasonally adjusted data rather than the cyclical-irregular movements should be smoothed.   If it is later desired to eliminate trend, the equation may be determined from the smoothed data.

A number of alternatives are open to the analyst in the choice of moving averages.   Probably the most commonly used is a simple 3-month moving average.   It frequently happens, on the other hand, that such a moving average introduces small inverse fluctuations into the series.   This difficulty can be overcome and a smoother curve obtained by the use of a 5-month average.   Such an average may, however, be too smooth; it may iron out turning points that are significant.   In the present study a binomially weighted 3-month moving average (central item weighted double) has been used in order to attain maximum sensitivity as well as smoothness.   The reader is already familiar with the computation of moving averages in general, and binomially weighted moving averages in particular (see pages 421–426); consequently, the mechanics of computation will not be further discussed.   Results are shown in column 5 of Table 128 and by the dotted line of Chart 199B.

In general it may be said that the moving average to be chosen depends on (1) the irregularity of the data (in respect to both amplitude and duration), and (2) the extent to which it is desired to smooth the data.   If the data are not very irregular, 3 months may be sufficient.   The more irregular are the data, the larger is the number of months required.   But the larger the number of months taken, the less flexible is the moving average For very irregular data the writers have found that a 5-month average weighted 1, 2, 4, 2, 1 gives a smooth, sensitive curve, without excessive

labor.   Note that the sum of the weights is 10, which eliminates a division.[1]

Comparison of cyclical movements.   One reason for wishing to isolate cyclical movements is that they may be compared with cyclical movements of other series.   Possibly it may be discovered that one series consistently precedes the other in its turning points, and thus the cyclical movements of the former may be used to forecast those of the latter.   Difficulty is experienced, however, in visualizing any time relationships between the cycles of the series if the amplitudes of their fluctuations differ appreciably. Thus, from Chart 200A it appears that pig iron production fluctuates about twice as violently as does United States magazine advertising.   Because of this fact the two curves do not lie close to each other throughout, and comparison of their movements is a little difficult.

A simple remedy for this difficulty is to use for magazine advertising a scale about twice as large as that chosen for pig iron production.   If a considerable degree of accuracy is required, a more satisfactory procedure is to adjust the two series so that they will have the same amplitude, and use only one scale for the two series.   The measure of amplitude usually chosen is the standard deviation.

The customary procedure for making this amplitude adjustment is illustrated in Table 130.   First, each series is converted into percentage deviations from normal, by subtracting 100 from each item.   Next, the standard deviation of each series is computed by the usual formula,

$$\sigma = \sqrt{\frac{\Sigma X^2}{N.} - \left(\frac{\Sigma X}{N}\right)^2}.$$

---

[1] For some purposes it may be desirable not merely to reduce minor irregularities, but as nearly as possible to completely eliminate all irregular movements, leaving only cycles.   For such purposes rather complicated moving averages are occasionally used Thus, the cyclical movements of Chart 143A were based upon a moving average weighted as follows: -1, -3, -5, -5, 2, 6, 18, 33, 47, 57, 60, 57, 47, 33, 18, 6, 2, -5, -5, -3, -1.   A characteristic of this particular weight pattern is that, if it is fitted to a second (or third) degree curve of simple polynomial series, it will fall exactly on that curve. Nevertheless, the results are not so smooth as might be desired, nor is it sufficiently flexible at the cyclical turning points.   (Smoothness is sometimes measured by taking the sum of the squares of the third differences.   The smaller the sum, the smoother is the series.)

The computation of this moving average is not so difficult as might be expected The procedure to be applied to deseasonalized data is as follows: Take a 5-month moving total of a 5-month moving total of a 7-month moving total.   Take a weighted 7-month moving total of the results with weights as follows: -1, 0, 1, 2, 1, 0, -1 Divide by 350 ($= 5 \times 5 \times 7 \times 2$).   The reader might at this point refer to pages 500–501, where a 43-term moving average for use with data that have not been deseasonalized was described.   That moving average removes seasonal as well as irregular movements.   Experimentation with the 43-term formula, however, leads the writers to the conclusion that it smooths out too much of the amplitude of cycles that are very short or have very sharp turning points, and sometimes does not coincide well with cyclical turning points.   For further discussion, see Frederick R. Macaulay, *The Smoothing of Time Series*, National Bureau of Economic Research, New York, 1931.

Even though the cycles have been put in deviation form, the deviations do not cancel out exactly; hence the correction factor in the above formula is necessary.   Finally each series is expressed in units of its standard
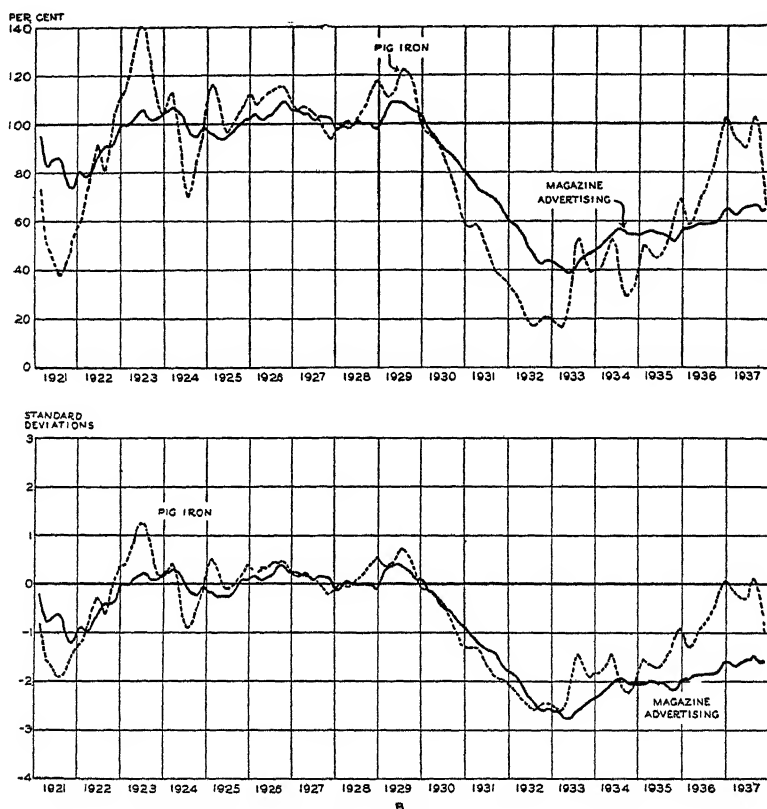
Chart 200.   Cyclical Movements of United States Magazine Advertising and of Pig Iron Production (A) as Percentages and (B) as Deviations in Units of Their Standard Deviations, 1921-1937.   (Each series has been smoothed by a binomially weighted 3-month moving average.   For magazine advertising data see Table 130; for source of pig iron production, original data, see Chart 139.)

deviation by dividing by its standard deviation.   In the present instance we find for United States magazine advertising:

$$\sigma = \sqrt{\frac{173,828.54}{202} - \left(\frac{3,885.8}{202}\right)^2} = 22.15.$$

This compares with a value for σ of 32.25 for pig iron production.   Therefore, when the cyclical deviations of magazine advertising are divided by 22.15 and pig iron by 32 25, the variations in the latter will be reduced a

relatively large amount. The amplitude of fluctuation of the series will be more nearly the same; furthermore, each now has the same degree of variability (that is, each has a standard deviation of unity).

Inspection of Chart 200B, which has a scale running from -4σ to $+3\sigma$,

TABLE 130

CALCULATION OF CYCLICAL DEVIATIONS OF UNITED STATES MAGAZINE ADVERTISING IN UNITS OF STANDARD DEVIATIONS, 1921-1937

Year and month (1)	Cyclical relatives (per cent) (2)	Deviations from 100 [Col. 2 − 100] (3)	Squared deviations (4)	Deviations in terms of σ [Col. 3 ÷ 22 15] (5)
1921:				
January..	. .		.	. .
February.	95 3	− 4.7	22 09	−0.21
March.	86 6	−13.4	179 56	− .60
April	83.4	−16 6	275 56	− .75
May	84.3	−15 7	246 49	− .71
June	85 6	−14.4	207 36	− .65
July	86 3	−13.7	187.69	− .62
August	85 0	−15 0	225 00	− .68
September	79 4	−20.6	424.36	− .93
October. .	74 3	−25.7	660 49	−1.16
November. . .	73.6	−26.4	696 96	−1.19
December . ..	76.8	−23.2	538 24	−1.05
1937:				
January.	64.8	−35 2	1,239.04	−1.59
February.........	62.9	−37.1	1,376 41	−1 67
March.	62.5	−37.5	1,406.25	−1 69
April	63 6	−36 4	1,324 96	−1.64
May..............	64.8	−35.2	1,239 04	−1.59
June	65 6	−34.4	1,183 36	−1.55
July	66 0	−34 0	1,156 00	−1.53
August.	66 6	−33.4	1,115 56	−1.51
September	65.6	−34.4	1,183 36	−1.55
October	64 4	−35.6	1,267 36	−1.61
November.. ..	65.1	−34.9	1,218 01	−1.58
December . . .	.	. .	...	...
Total .	. .	−3,885.8	173,828.54	...

Source· Table 128

verifies the fact that each series now has about the same amplitude. This type of chart is frequently referred to as a *cycle* chart, since its object is to facilitate comparison of cycles. It is easily observable from this chart that the turning points of pig iron typically occur before those of magazine advertising. The chart, of course, does not imply that pig iron produc-

tion results in magazine advertising; in itself the chart gives no clue to the causal relationship.

Direct Method

The residual method is rather laborious if the sole object of the analysis is to isolate cycles. A simpler and more direct method is desirable, but unfortunately no direct method has yet been devised which accomplishes this result with any great degree of perfection.

For rough and ready reference, Chart 201, on which the unadjusted
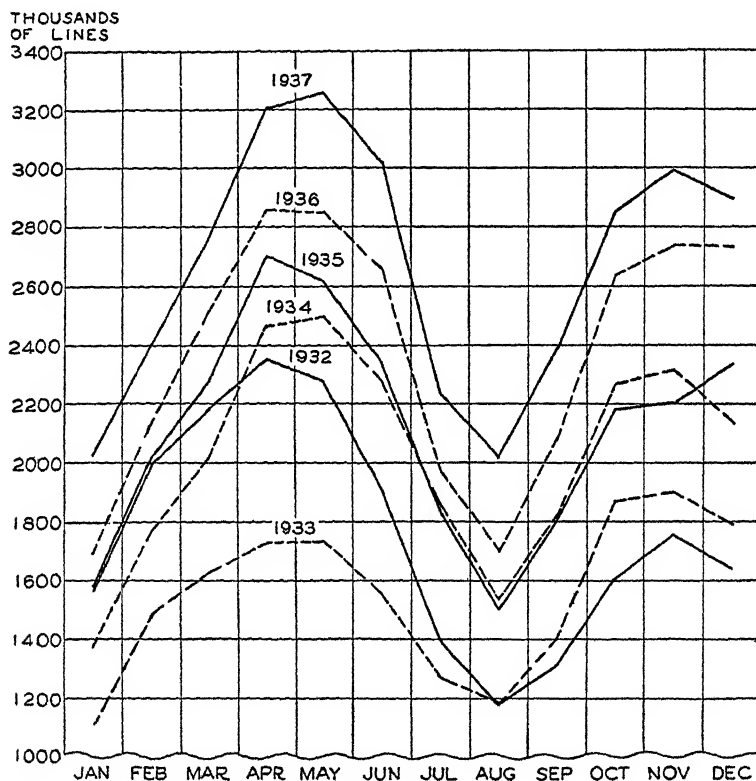
Chart 201. United States Magazine Advertising, by Months, 1932–1937. (Data of Table 109, Col. 2.)

data have been plotted, has much to commend it. Although little idea of the trend can be obtained (since if any considerable number of years were included, the crossing and re-crossing of the lines would be very confusing), the seasonal movement is unmistakable. More important for our purposes, a rough idea of the cycles can be obtained by comparing the

level and slope of any line with those preceding and following it in point of time.   Thus the latter half of 1932 and the first half of 1933 appear as depression periods, and there appears also a minor recession in the middle of 1935.

But a general purpose chart, such as Chart 201, though it gives a rough idea of several factors, fails to give a very precise idea of any one.   A more exact comparison can be made by the use of a little simple arithmetic. Thus each month might be expressed as a percentage of the corresponding month in the preceding year.[2]   The procedure, of course, is to divide each January value by the January value for the year before; likewise for February; and so on.   This procedure roughly eliminates seasonal variation and secular trend, though the general level of the percentages will be above 100 if the trend is upward, and below 100 if the trend is downward.   The results of this procedure are shown in section A of Chart 202.   The "cycles" are thrown into clear relief, but they are not the same sort of fluctuations with which we are familiar.   They represent not the cyclical level, but the cyclical change.   Thus the 1934 value is very high, though the usual cyclical analysis (as shown by section B of this chart) indicates that in every month of that year advertising was at least 40 per cent below normal. The explanation is that 1933 was a year of extreme depression, and 1934 was high *in comparison*!   Although the same movements may be detected in sections A and B of this chart, they take an unusual form in section A, and the mental readjustment which must be made in interpreting the cycles of this section makes the procedure of doubtful utility to most persons and for most purposes.   Another defect of this procedure is that irregularities in the data are magnified.   For instance, the conjuncture of a minor irregular high in the given month and a minor irregular low in the corresponding month of the preceding year will produce rather a large positive variation in the derived series for the given month.

A variation of this method which gives improved results is to express the data as a percentage of the average of the corresponding month for several of the preceding years.   In section C of Chart 202, the three preceding years are used.   The number should coincide with the average length of the cycle in the series under consideration.   As can be seen from the chart, this method gives results more like those of the orthodox method of section B than does the method shown in section A.   Probably the chief objection to this method is that cycles are not uniform in duration or amplitude, and rather serious distortion of the data still results.   There-

---

[2] This method was devised by M. A. Brumbaugh.   See his *Direct Method of Determining Cyclical Fluctuations of Economic Data*   Prentice-Hall, Inc., New York, 1926. Brumbaugh also adjusts for the small residue of trend which is left after the first series of divisions.

fore, if accurate results are required, the residual method is to be recommended in preference to the direct method.

Harmonic Analysis

After obtaining cyclical-irregular movements, an alternative method to smoothing by a moving average, or in addition to such a procedure, is to
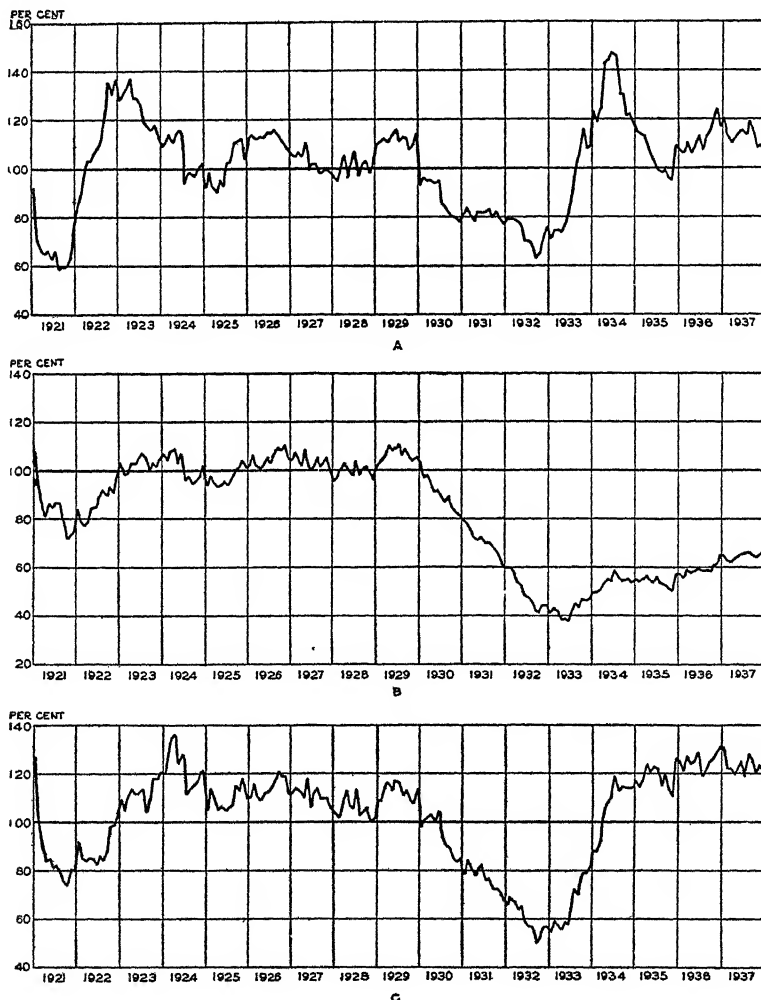
Chart 202.   United States Magazine Advertising Cyclical-Irregular Movements Computed by Three Methods, 1921-1937.   A. Per Cent of Corresponding Month in Preceding Year; B. Per Cent of "Normal"; C. Per Cent of Average of Corresponding Month in Three Preceding Years.   (For original data see Table 109, Col. 2; for cyclical-irregular movements by residual method see Table 128.)

fit a mathematical curve to the data.   The simple procedure here will
seldom be found appropriate, since cycles seldom exhibit a simple period-
icity, but the method is to be regarded as an introduction to more complex
methods of the same general nature.   Non-acetate rayon deliveries, how-
ever, seem to exhibit a fairly regular 2-year cycle, and hence these data
were selected for purposes of illustration.   According to Stanley B. Hunt,
editor of *Rayon Organon*, "the chief cause of this textile production cycle
apparently is a periodic fluctuation in the stocks or inventories of goods
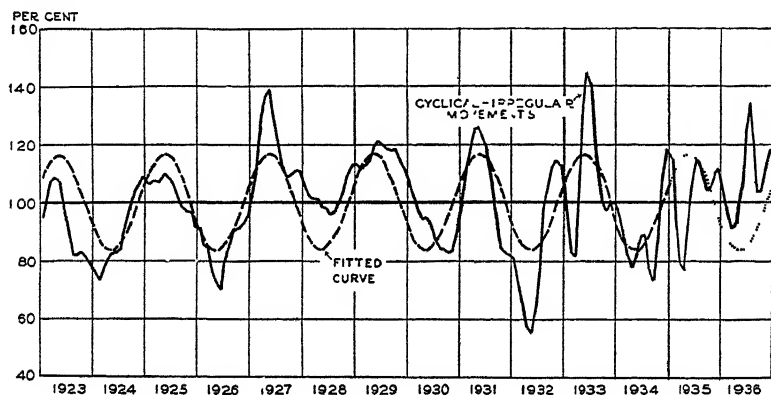
Chart 203.   Cyclical Movements of Non-Acetate Rayon Deliveries, and Sine-Cosine
Curve.   (Data of Table 131 and 133.)

held at all stages of production and distribution."[3]   These data were ad-
justed for trend and seasonal, and, on account of violently irregular move-
ments, were partially smoothed by a 5-month moving average, weighted
1, 2, 4, 2, 1.   These data are shown in Table 131 and Chart 203 (solid line).

The procedure falls into two steps.   (1)   By means of *periodogram* an-
alysis an attempt is made to discover the periodicity of the data and its
average cyclical pattern.   (2)   A *periodic curve* is fitted to the average
pattern and applied to the series being analyzed.

Periodogram analysis.   Our first objective is to discover the periodicity
of the rayon data.   Let us assume that the periodicity is 24 months.   In
Table 132 the data of Table 131 through December 1934 are arranged in
rows of 24 successive items.   (December 1934 was arbitrarily chosen as a
terminating point for this illustration since the data in the following years
do not conform so well to the general 2-year pattern).   The average of
each column is now taken.[4]   The highest average is 122.5, and is for the

---

[3] See Textile Economic Bureau, Inc., *Rayon Organon*, December 8, 1937, p 171.

[4] The average selected was the arithmetic mean, although a median or modified
mean might be used.

TABLE 131

CYCLICAL MOVEMENTS OF NON-ACETATE RAYON DELIVERIES, 1923–1936

(Cyclical-irregular movements smoothed by moving average, weighted 1, 2, 4, 2, 1)

Year	January	February	March	April	May	June	July	August	September	October	November	December
1923	95.0*	101.0*	107.7	108.9	107 1	97 1	89 7	81 8	82 2	83.1	80 8	78 9
1924	75.8	73.4	76 1	80 6	82 8	83 0	84 0	91 2	99.2	103 7	105.9	109 0
1925	106 7	107 5	107.6	107 6	110 4	108 6	106.6	102 4	98 8	96.8	97 3	91.5
1926	91.7	85.7	83.8	76 4	73 2	70 1	80.1	86 6	90 9	91.5	92 7	96 0
1927	101.9	112.0	126.3	137.3	138 9	129.2	120 4	111.8	108 9	109.6	110.8	111.0
1928	106.0	102.7	101.6	101.5	99 1	97.8	96 0	96 9	100 2	105.3	109.4	113 1
1929	112.9	111.0	113.2	114.6	119.2	121.0	120.1	118.6	117 9	118.3	114 5	11ı 2
1930	107.1	102 6	98.4	94.6	95 0	92.5	86.8	83.8	84 0	82 8	82 9	88.6
1931	98.2	109.7	117.3	125.0	125.9	122 2	118 0	106 2	92.7	84.1	82 8	82.4
1932	81.2	74.1	67.3	58.0	55.5	62 6	80 2	98 7	106 0	112.2	114 1	112 7
1933	97.4	82.8	81.5	104 6	133 3	144.4	140 4	116 6	100.9	97.4	100 0	99.9
1934	97.3	90.7	82.7	78 1	81.9	88.4	88 8	78 1	73 4	85.8	103 6	118.6
1935	114 7	95.0	79 2	76.9	93.0	106.3	114 9	110 9	104 5	105 5	109.8	112.0
1936	105.4	97.7	90.8	93.5	106 1	124 9	134 2	120 0	104 0	104 2	113 7	118 3

* Partly estimated

TABLE 132

PERIODOGRAM ANALYSIS OF RAYON CYCLES $(T = 24)$

Measure	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	Total
	95.0	101 0	107 0	108 9	107 1	97 1	89 7	81 8	82 2	83 1	80 8	78 9	75 8	73 4	76 1	80 6	82 8	83 0	84 0	91 2	99 2	103 7	105 7	109 0	
	106 7	107.5	107 6	107 6	110.4	108 6	106 6	102 4	98 8	90 8	97 3	91 5	91 7	85 7	83 8	76 4	73 2	70 1	80 1	86 6	90 9	91 5	92 7	96 0	
	101 9	112 0	126 3	137 3	138 9	129 2	120 4	111 8	108 9	109 6	110 8	111 0	106 0	102 7	101 6	101 5	99 1	97 8	96 0	96 9	100 2	105 3	109 4	113 1	
	112 9	111 0	113.2	114 6	119 2	121 0	120 1	118 6	117 9	118 3	114 5	111 2	107 1	102 6	98 4	94 6	95 0	92 5	86 8	83 8	84 0	82 8	82 9	88 6	
	98 2	109 7	117 3	125 0	125 9	122 2	118 0	106 2	92 7	84 1	82 8	82 4	81 2	74 1	67 3	58 0	55 5	62 6	80 2	98 7	106 0	112 2	114 1	112 7	
	97 4	82 8	81 5	104 6	133.2	144 4	140 4	116 6	100 9	97 4	100 0	99 9	97 3	90 7	82 7	78 1	81 9	88 4	88 8	78 1	73 4	85 1	103 6	118 6	
Average pattern $\bar{Y}$..	102.0	104 0	108 9	116 3	122 5 (H)	120.4	115.9	106 2	100.2	98 2	97 7	95.8	93 2	88 2	85 0	81 5	81 2 (L)	82 4	86 0	89 2	92 3	96 9	101 4	106 4	2,371 9
Adjusted pattern* $\bar{Y}'$	103 2	105 2	110 2	117 7	124 0	121 8	117 3	107 5	101 4	99 4	98 0	96 9	94 3	89 2	86 0	82 5	82 2	83 4	87 0	90 3	93 4	98 0	102 6	107 6	2,400 0

* Average values multiplied by 1 011847 = 2,400.0 ÷ 2,371 9.
H = high. L = low.
Source: Table 131.

557

fifth month; whereas the lowest average, 81.2, is found for the seventeenth month. The range from high month to low month is 122.5 − 81.2 = 41 3.

We now repeat this procedure with an arrangement of 23 successive items in a row. The table is not reproduced here, but the reader can easily verify that the last number in the first row would be 105.9, which refers to November 1924, and that the final item of the entire series (23rd item of row 6) would be 88.4, the cyclical relative for June 1934. Actual computation gives a range of means of 27.5. Repeating again with a 25-month period, running through June 1935, we find a range of 37.7.
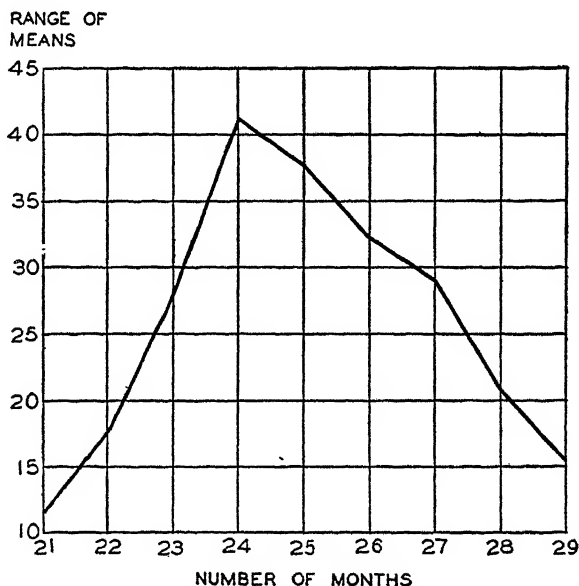
Chart 204. Periodogram of Cyclical Movements of Non-Acetate Rayon Deliveries, 1923-1934. (For data see below.)

This procedure, with different assumptions as to periodicity, must be repeated until the statistician is satisfied that he has discovered the true periodicity. This is taken as the periodicity that gives the greatest range of column means. For nine different trials the results are as follows:

Assumed periodicity (months)	Range of column means
21	11.6
22	17.7
23	27.9
24	41 3
25	37 7
26	32.1
27	29.0
28	20.7
29	15.4

These results are shown graphically in Chart 204, known as a *periodogram*.

The column means of Table 132 supply us with numerical values for the average cyclical behavior of rayon consumption. Following the procedure adopted in the computation of seasonal indexes, however, we shall adjust the values to average 100 per cent. These adjusted data are in the last row of Table 132, and are shown by the solid line of Chart 205.
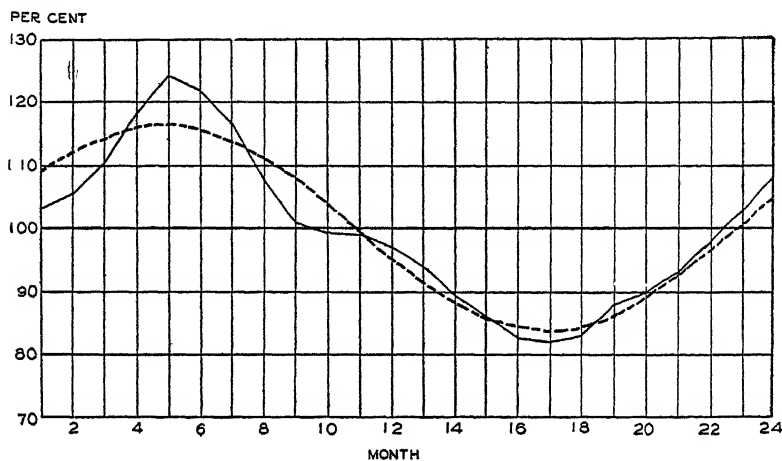
Chart 205. Average Cyclical Pattern of Non-Acetate Rayon Deliveries, 1923–1934 (Solid Line) and Sine-Cosine Curve (Dashed Line). (Data of Table 133.)

Fitting a periodic curve. We shall now fit a sine-cosine curve to the adjusted cyclical pattern ($\overline{Y}'$ values) The curve type is

$$Y_C = \overline{Y}' + A \sin \left(\frac{360}{T} X\right)^{\circ} + B \cos \left(\frac{360}{T} X\right)^{\circ},$$

where

$$T = \text{the periodicity in months,}$$

$$A = \frac{2}{T} \Sigma \left[\overline{Y}' \sin \left(\frac{360}{T} X\right)^{\circ} \right],$$

$$B = \frac{2}{T} \Sigma \left[\overline{Y}' \cos \left(\frac{360}{T} X\right)^{\circ} \right].$$

A further observation concerning this equation is that the range of the fitted curve from peak to trough is $2 \sqrt{A^2 + B^2}$.

Since T has been found to be 24, and $\overline{Y}' = 100$, we may immediately write:

$$Y_C = 100 + A \sin (15X)^{\circ} + B \cos (15X)^{\circ},$$

$$A = \frac{\Sigma[\overline{Y}' \sin (15X)^{\circ}]}{12},$$

$$B = \frac{\Sigma[\overline{Y}' \cos (15X)^{\circ}]}{12}.$$

Computations for fitting this curve are found in Table 133. Note that the first value of X is taken as 1 instead of the usual 0. Sines and cosines for columns 3 and 4 are read from Appendix L. The $\overline{Y}'$ values are the adjusted values in the last row of Table 132. Substituting in the formulae above, we find:

$$A = \frac{186.23}{12} = 15.519,$$

$$B = \frac{58.96}{12} = 4.9133.$$

The equation, then, is $Y_C = 100 + 15.519 \sin (15X)° + 4.9133 \cos (15X)°$. The last three columns of the table are self-explanatory. The highest computed value, 116.3, we find in the fifth month, and the lowest, 83.7, in the seventeenth month. The range, therefore, is $116.3 - 83.7 = 32.6$. This checks with $2\sqrt{A^2 + B^2} = 2\sqrt{(15.519)^2 + (4.9133)^2} = 2 \times 16.3 = 32.6$.

The computed values of column 10 are shown by the broken line of Chart 205. The fit is good from the eleventh month on, though much is to be desired in the early part of the curve. The same curve is shown by the broken line of Chart 203. Since sin 375° is the same as sin 15°, and cos 375° is the same as cos 15°, Y_C is the same for $X = 25$ as for X is 1 Likewise, Y_C when $X = 26$ is the same as when $X = 2$, and so on. The fitted curve therefore repeats itself each 24 months. As can be seen, the sine-cosine curve fits reasonably well from 1923 through 1934, but the extension through 1937 is unsatisfactory. As is so frequently the case, generalizations which are valid historically do not work when extended into the future. The reason is that conditions change and causes which were important in the past give way to newer and more potent causes. Again it might be noticed that time series since about 1933 have become less regular in their behavior. Another difficulty with the simple, though somewhat laborious, procedure which has been illustrated, is that the amplitude of the fitted curve is smaller than the amplitude of the original data. This is because the length of each cycle is not exactly the same, and therefore the column of the periodogram table which contains the high mean does not contain all the cyclical highs, nor does the column with the low mean contain all the cyclical lows. For this reason the average pattern (to which the periodic curve is fitted) has smaller amplitude than the original data.

Cyclical Averages

The method of harmonic analysis which has been discussed assumes: (1) that cycles are a variety of periodic movement; (2) that they are similar in pattern; (3) that the pattern can be described by a mathematical

TABLE 133

PERIODIC CURVE FITTED TO NON-ACETATE RAYON CYCLES ($T = 24$)

X (1)	15X (2)	sin (15X)° (3)	cos (15X)° (4)	Ȳ' (5)	Ȳ' sin (15X)° (6)	Ȳ' cos (15X)° (7)	Computation of cycle values		
							A sin (15X)° (8)	B cos (15X)° (9)	Yc [100 + Col. 8 + Col. 9] (10)
1	15	.2588	.9656	103.2	26.71	99.65	4.16	4.74	108.9
2	30	.5000	.8660	105.2	52.60	91.10	7.76	4.25	112.0
3	45	.7071	.7071	110.2	77.92	77.92	10.97	3.47	114.4
4	60	.8660	.5000	117.7	101.93	58.85	13.44	2.46	115.9
5	75	.9656	.2588	124.0	119.73	32.09	14.99	1.27	116.3
6	90	1.0000	.0000	121.80	121.80	0	15.52	0	115.5
7	105	.9656	—.2588	117.3	113.26	—30.36	14.99	—1.27	113.7
8	120	.8660	—.5000	107.5	93.10	—53.75	13.44	—2.46	111.0
9	135	.7071	—.7071	101.4	71.70	—71.70	10.97	—3.47	107.5
10	150	.5000	—.8660	99.4	49.70	—86.08	7.76	—4.25	103.5
11	165	.2588	—.9656	98.9	25.60	—95.50	4.16	—4.74	99.4
12	180	.0000	—1.0000	96.9	0	—96.90	0	—4.91	95.1
13	195	—.2588	—.9656	94.3	—24.40	—91.06	—4.16	—4.74	91.1
14	210	—.5000	—.8660	89.2	—44.60	—77.25	—7.76	—4.25	88.0
15	225	—.7071	—.7071	86.0	—60.81	—60.81	—10.97	—3.47	85.6
16	240	—.8660	—.5000	82.5	—71.44	—41.25	—13.44	—2.46	84.1
17	255	—.9656	—.2588	82.2	—79.37	—21.27	—14.99	—1.27	83.7
18	270	—1.0000	.0000	83.4	—83.40	0	—15.52	0	84.5
19	285	—.9656	.2588	87.0	—84.01	22.52	—14.99	1.27	86.3
20	300	—.8660	.5000	90.3	—78.20	45.15	—13.44	2.46	89.0
21	315	—.7071	.7071	93.4	—66.04	66.04	—10.97	3.47	92.5
22	330	—.5000	.8660	98.0	—49.00	84.87	—7.76	4.25	96.5
23	345	—.2588	.9656	102.6	—26.55	99.10	—4.16	4.74	100.6
24	360	—.0000	1.0000	107.6	0	107.60	0	4.91	104.9
Total				2,400.0	186.23	58.96			2,400.0

561

equation. In practice it is found that most economic series are not peri-odic, and that it is difficult adequately to describe them by mathematical curves. Wesley C. Mitchell, in studying business cycles, has come to the conclusion that different cycles of a given series are sufficiently alike in pattern to justify averaging them together and making a number of meas-urements of average behavior. Although Mitchell's method is not widely used, the importance of his extensive studies justifies a brief description of some of his basic procedures.[5] Computations are carried out by the National Bureau of Economic Research under Mitchell's direction.

As a preliminary step Mitchell adjusts the data for seasonal variation but not for trend within cycles, since he believes the study of combined cycle and intra-cycle trend to be useful. Using the deseasonalized data, average patterns are obtained for *specific cycles* and for *reference cycles*, as will be described in turn.

Specific cycle analysis. Although no adjustment is made for intra-cycle trend, the data are adjusted for trend between cycles (inter-cycle trend) before averaging the different cycles. This is done by expressing each individual month as a percentage of the average for that cycle. In order to do this, it is necessary to break the series into specific cycles each run-ning from low to low. No objective procedure is adopted for accomplish-ing this, but turning points are selected largely by inspection of a chart of deseasonalized data, such as Chart 206. On this chart the *peaks* and *troughs* selected for the specific cycles of United States magazine adver-tising are marked by asterisks. The dates selected are given as follows. By definition, *revival* occurs in the month following the cyclical trough; similarly, *recession* refers to the month following a cyclical peak.

DATES OF SPECIFIC CYCLES IN UNITED STATES MAGAZINE ADVERTISING

Cycle	Initial revival	Peak	Trough	Terminal revival
1	December 1918	August 1920	October 1921	November 1921
2	November 1921	April 1924	September 1924	October 1924
3	October 1924	November 1926	January 1928	February 1928
4	February 1928	July 1929	June 1933	July 1933

Now the average value of the deseasonalized data for each cycle *from revival through trough is obtained.*

Cycle	Total of deseasonalized data (thousands of lines)	Duration of cycle (months)	Average value for cycle
1	80,037	35	2,287
2	81,703	35	2,334
3	113,945	40	2,849
4	168,781	65	2,597

[5] See also Wesley C. Mitchell and Arthur F. Burns, *The National Bureau's Measures of Cyclical Behavior,* Bulletin 57, July 1, 1935, of the National Bureau of Economic Research, New York.
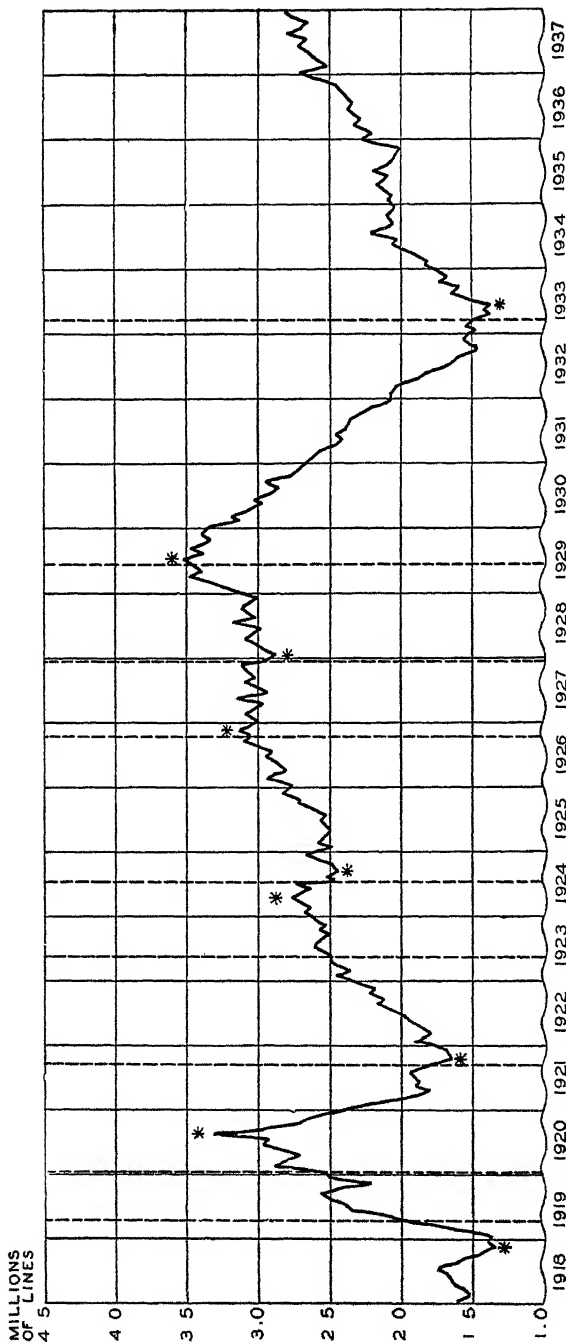
Chart 206. United States Magazine Advertising Adjusted for Seasonal Variation and Dates of Peaks and Troughs of Specific Cycles and Reference Cycles, 1918-1933. (Specific cycle turning points are shown by asterisks; those for reference cycles by vertical dotted lines. For deseasonalized data, 1921-1937, see Table 128. For dates of peaks and troughs see p. 562 and p. 568.)

Next the deseasonalized data are divided by the average cycle value for each cycle and multiplied by 100. The percentages so obtained include for each cycle a span from the month preceding initial revival (that is, the trough) through the month following the terminal revival.[6] For instance, the percentages for the first cycle run from November 1918 through December 1921, and for the second cycle from October 1921 through No-

STAGES OF CYCLE 1, UNITED STATES MAGAZINE ADVERTISING

Stage number	Name of stage	First month	Month on which centered	Last month	Duration in months*
I	Initial revival .	,	Dec 1918		1
	Period of expansion:				
II	First third . . .	Jan. 1919	Mar.–Apr 1919	June 1919	6
III	Second third .	July 1919	Oct 1919	Jan 1920	7
IV	Last third . .	Feb 1920	May 1920	Aug 1920	7
V	Recession		Sept. 1920	. .	1
	Period of contraction:				
VI	First third . . .	Oct. 1920	Nov –Dec. 1920	Jan. 1921	4
VII	Second third	Feb 1921	Mar –Apr. 1921	May 1921	4
VIII	Last third	June 1921	Aug. 1921	Oct 1921	5
IX	Terminal revival ..	.	Nov. 1921		1

* Whenever the months in a period are not divisible by three, the adjustment is made in the middle stage For this purpose, initial revival is considered to be a part of the first third of expansion, and recession a part of the first third of contraction.

vember 1924. It is to be noticed that October, November, and December, 1921, are included as the last three months of cycle 1 and the first three months of cycle 2. Although there is an overlapping of months between cycles, the percentages for these months are different in the two cycles because they are based on different averages. Since these data are basic, they are shown as Table 134.

After inter-cycle trend is eliminated as described, the different cycles are further made comparable by dividing each into nine stages. The first stage is called *initial revival,* and is the month following the initial trough. The next three stages are equal thirds of the period of *expansion,* which runs from the month following that of initial revival through the peak. The fifth stage, that of *recession,* is the month following the peak. The

---

[6] These percentages are necessary in order to obtain averages for the nine stages which are defined in the following discussion. Possibly it might have been better, here and later, to have had stages I and IX represent the troughs instead of the revival months and stage V the peak instead of the recession month; but in order to avoid confusion, we are following Mitchell's procedure.

TABLE 134

UNITED STATES MAGAZINE ADVERTISING SPECIFIC CYCLES, NOVEMBER 1918–AUGUST 1933

(Seasonally adjusted data expressed as per cent of average for each cycle)

Year	January	February	March	April	May	June	July	August	September	October	November	December
Cycle 1:												
1918											58.6	60.3
1919	60.0	66.0	77.1	86.8	92.9	102.6	105.0	109.7	112.3	106.2	97.1	109.7
1920	111.4	126.3	122.9	119.0	125.5	130.0	128.6	144.7	129.0	118.3	115.6	108.6
1921	102.1	89.1	82.3	77.7	82.6	82.4	84.3	85.1	77.7	71.3	72.9	74.3
Cycle 2:												
1921										69.8	71.5	72.8
1922	82.5	77.9	77.2	78.7	83.6	85.4	89.5	93.2	91.1	94.5	92.6	99.3
1923	105.7	101.4	102.8	107.4	107.6	109.9	112.9	110.9	107.0	110.0	108.9	111.8
1924	115.2	112.7	117.2	118.6	112.8	118.1	105.9	108.2	105.3	106.4	108.9	
Cycle 3.												
1924									86.3	87.2	89.2	93.8
1925	87.4	90.7	89.3	87.7	88.3	90.1	89.0	91.6	95.7	96.0	99.8	97.8
1926	98.7	103.1	99.8	99.3	100.1	103.5	101.9	106.3	109.0	108.1	110.3	106.0
1927	105.0	108.5	107.2	103.9	110.6	103.1	104.1	108.7	106.3	107.0	109.9	104.9
1928	101.3	102.8	105.8									
Cycle 4:												
1928	111.1	112.8	116.0	119.7	116.5	114.9	121.9	115.4	118.5	120.4	117.7	114.9
1929	121.9	125.1	130.1	133.0	131.5	133.1	135.2	130.1	133.5	129.4	129.2	131.0
1930	128.5	121.0	122.4	120.8	114.9	116.3	113.3	110.4	114.1	107.8	106.0	105.1
1931	102.6	101.6	98.9	94.3	93.8	94.7	92.4	92.0	90.8	88.5	85.2	80.9
1932	80.5	80.1	78.3	73.2	71.9	66.0	64.4	63.0	57.5	57.3	61.1	61.2
1933	57.2	59.7	58.4	53.9	54.7	53.5	58.8	63.6				

Source: 1921–1933, derived from Table 128   For source of original data, see Table 85

next three stages are equal thirds of the period of *contraction*, which runs from the month following that of recession through the trough. The last stage is called *terminal revival*, and is the month following the terminal trough. In order to clarify these statements, we show, in tabular form, on page 564, the nine stages of the first specific cycle of magazine advertising.

Note that, although the number of months in the last column totals 36, the duration of the cycle, from December 1918 through October 1921, is only 35 months. This is a formal discrepancy due to the overlapping of the stage of terminal revival of one cycle with the period of initial revival of the next.

A standing or average value for each stage of each cycle is now computed from the data of Table 134. Standings for stages I, V, and IX are taken as the average of the three months centering respectively on the months of initial revival, recession, and terminal revival. This is done in order to obtain more representative values for these stages. The standings for the different stages of the first cycle are found to be as follows:

CYCLICAL PATTERN OF CYCLE 1

Stage	Standing
I	59 6
II	80.9
III	107.3
IV	128 1
V	130 7
VI	111.2
VII	82 9
VIII	80 2
IX	72 8

Since the pattern of each cycle is not exactly the same, an average, for all cycles, of stage I is obtained, and of stage II, and so on. The data and results are shown in Table 135. The averages of this table constitute the average pattern of specific cycles in United States magazine advertising. As a very rough indication of the reliability of these averages, average deviations are shown at the bottom of the table.

Reference cycle analysis. The object of reference cycle analysis is to determine how a specific series behaved, on the average, during cycles in general business. The analysis is carried through in precisely the same fashion as for specific cycles, except that the dates chosen for revivals, peaks, and troughs are those of general business cycles rather than those of the specific series being analyzed. These reference cycle dates are established subjectively after examination of the material in Thorp's *Business Annals* (National Bureau of Economic Research, New York, 1926) and

TABLE 135

COMPUTATION OF SPECIFIC CYCLE PATTERN OF UNITED STATES MAGAZINE ADVERTISING, 1918–1933

Cycle	Initial revival (I)	Expansion			Recession (V)	Contraction			Terminal revival (IX)
		First third (II)	Second third (III)	Last third (IV)		First third (VI)	Second third (VII)	Last third (VIII)	
1	59.6	80.9	107 3	128.1	130 7	111.2	82 9	80 2	72.8
2	71.4	82.3	101 2	112.5	116 5	118 1	105.9	106 8	106.9
3	87.6	89 6	96 5	104 3	107 1	106.2	106 6	105 9	103.3
4	113.3	117.8	118 1	131 3	132 9	119.9	92.1	62 0	58.6
Average..........	83.0	92.6	105 8	119 0	121 8	113 8	96 9	88 7	85 4
Average deviation .	17.5	12 6	6 9	10 6	10 0	5 2	9.4	17.6	19.7

Source: Derived from Table 134.

567

various statistical series. Post-war reference cycle dates, as established by Mitchell, are as follows:

Cycle	Revival	Peak	Trough	Duration (months)
1	May 1919	January 1920	September 1921	29
2	October 1921	May 1923	July 1924	34
3	August 1924	October 1926	December 1927	41
4	January 1928	June 1929	March 1933	63

Analysis along the lines indicated gives the following results:

Stage	Average value	Average deviation
I	90 0	10 2
II	94 6	11 0
III	101 0	8 2
IV	108 4	8 8
V	115 8	6 9
VI	114.9	5 4
VII	104.3	5 6
VIII	91.4	20 6
IX	83.9	21 9

Comparison of reference and specific cycles. From the measures already illustrated or described a number of interesting comparisons can be made. Many of these may most easily be seen by study of a chart, such as Chart 207. The horizontal scale represents time, while the vertical is per cent of average. First, notice that solid lines refer to the reference cycle pattern, while broken lines refer to the specific cycle pattern. If the turning point of each specific cycle coincided with the reference cycle dates, the two curves would be exactly the same. Variation between the two sets of dates produces two results: (1) variation in the pattern of the two curves; (2) smaller amplitude in the reference cycle pattern. In the present instance the conformity between the two series is high. (Among other measures that Mitchell makes are indexes of conformity, which will not be described here.) This indicates that magazine advertising is closely related to general business activity, as cause or effect, or in some other fashion.

The chart also indicates that magazine advertising tends to lag behind general business in its turning points. The facts concerning lag or lead of specific cycles with respect to the reference cycle are obtained by comparing the dates shown on page 562 and above, and summarized on page 569. Thus in Chart 207 the observation representing the first stage (initial revival) of the specific cycle is placed one-fourth of a month to the left of that representing the same stage of the reference cycle, while the
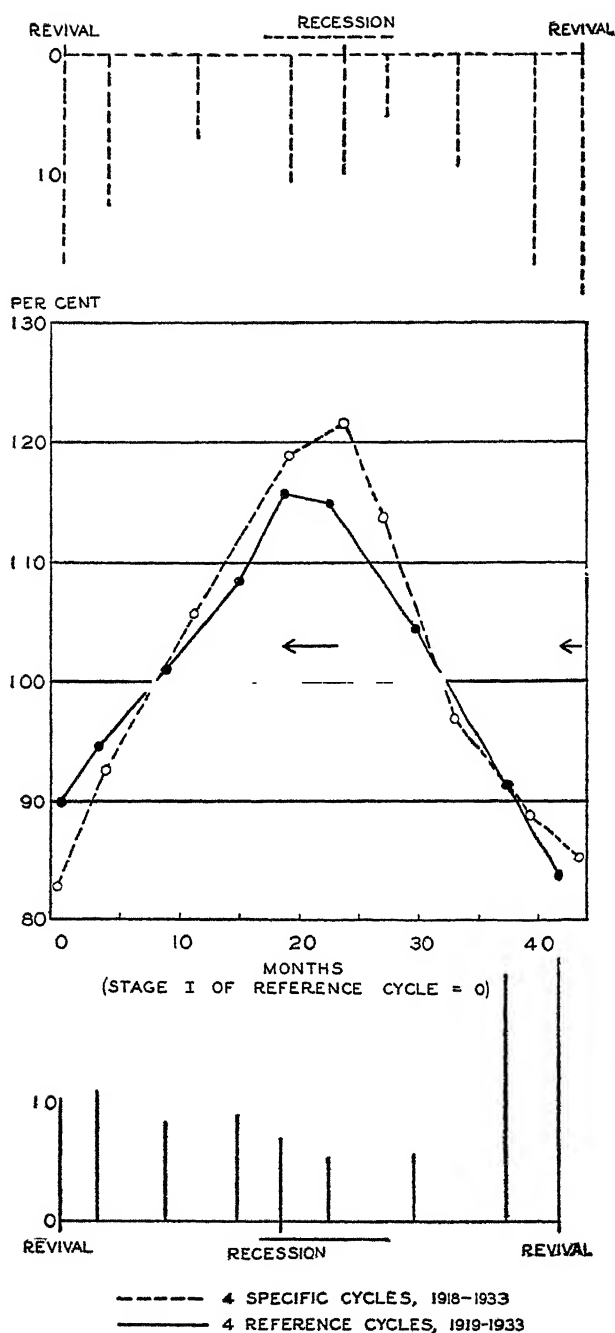
Chart 207. Cyclical Pattern of United States Magazine Advertising, 1918–1933.
For specific cycle pattern see Table 135; for further explanation read pp. 568–570.

fifth stage (recession) and the ninth stage (terminal revival) show the reference cycle further to the left by 5 and 1.75 months respectively. Not much reliance can be placed on the lead of magazine advertising at initial revival, however, since the negative average is due entirely to the lead of advertising one time out of four. The other two averages are probably significant; this belief is indicated by the two arrows pointing left on the chart. The length of the arrows indicates the number of months' lead, and the direction that the arrows point tells which leads (the specific cycle if the arrow points right; general business if the arrow points left).

Another interesting feature of this chart is the vertical lines at the top and bottom of the chart. These are on the same scale as the main part

LEAD (−) OR LAG (+) OF SPECIFIC CYCLE AT TURNING
POINTS OF REFERENCE CYCLE, IN MONTHS

(When the average is negative, the standing of the specific cycle at revival or recession is plotted to the left of that reference cycle stage in Chart 207, when it is positive, the specific cycle standing is plotted to the right)

Cycle	At initial reference trough	At reference peak	At terminal reference trough
1	−5	+7	+1
2	+1	+11	+2
3	+2	+1	+1
4	+1	+1	+3
Average	−0.25	5.00	1.75

of the chart, and indicate the average deviation of the standings for the different stages. These average deviations and the averages to which they refer are placed in a position along the horizontal scale proportional to the time elapsed from the center of one stage to the center of another. Finally it should be noted that the duration of the cycle is indicated by the length of the long horizontal line above the chart in the case of the specific cycle, and below it in the case of the reference cycle. These horizontal lines form the base lines with which the vertical lines are connected. The shorter horizontal lines paralleling the ones described refer to the average deviation of the duration of the cycles.

In connection with this type of analysis, still other measures of cyclical behavior are computed at the National Bureau of Economic Research. For instance, there are measures of cyclical amplitude, of percentage growth between cycles, of change from stage to stage. Space does not permit a discussion of these measures. but the reader is referred to the National

Bureau of Economic Research Bulletin 57 (July 1, 1935), *The National Bureau's Measures of Cyclical Behavior,* by Wesley C. Mitchell and Arthur F. Burns; and to mimeographed editions of the first three chapters of Mitchell's forthcoming book, *Business Cycles,* Volume II, *Analysis of Cyclical Behavior.*

Several methods of cycle analysis have been presented in this chapter. They each have strong points and weak points. The orthodox method, which isolates cycles by means of adjusting for all other types of movements, is probably on the whole the most satisfactory. Direct methods, though less laborious than this method, are difficult to interpret. The fitting of periodic curves is an attempt to generalize concerning cyclical behavior, but the regularities which it attempts to measure are not usually existent. The method of averages also attempts to generalize concerning cycles. Although it does not assume the same degree or kind of regularity, nevertheless it is not completely satisfactory. We are not certain that differences among cycles over a period of time are largely accidental. Also, the average deviations of Chart 207 are rather large, and, with only four cycles as observations, the averages plotted could not under any circumstances be considered very reliable. To the extent that cyclical behavior gradually changes, however, it might be possible to modify the method so as to give some idea of such trend.

Selected References

O. W. Blackett and W. P. Wilson: *A Method of Isolating Sinusoidal Components in Economic Time Series;* University of Michigan Press, Ann Arbor, 1938.

E. C. Bratt: *Business Cycles and Forecasting;* Business Publications, Inc., Chicago, 1937.

W. L. Crum, A. C. Patton, and A. R. Tebbutt: *Introduction to Economic Statistics,* pages 345–352; McGraw-Hill Book Co., New York, 1938. Illustrates the analysis of quarterly data.

H. T. Davis and W. F. C. Nelson: *Elements of Statistics* (Second Edition), pages 137–144; Principia Press, Bloomington, Indiana, 1937. A discussion of harmonic analysis.

F. R. Macaulay *The Smoothing of Time Series;* National Bureau of Economic Research, New York, 1931. Use of weighted moving averages. A readable discussion of a complex topic.

W. C. Mitchell: *Business Cycles, The Problem and Its Setting;* National Bureau of Economic Research, New York, 1928. Reading of the entire book is recommended. The material most closely paralleling Chapter XIX is on pages 249–262.

W. C. Mitchell, assisted by Arthur F. Burns: Forthcoming volume, *Business Cycles, Volume II, Analysis of Cyclical Behavior;* National Bureau of Economic Research. A preliminary draft of Chapters I and II was issued in mimeograph form in 1935, and of Chapter III in 1936. These chapters fully illustrate the statistical methods followed by Mitchell.

H. L. Rietz, Editor: *Handbook of Mathematical Statistics*, Chapter XI; Houghton Mifflin Co., Boston, 1924. An explanation of the periodogram and harmonic analysis.

J. R. Rigglemann and I. N. Frisbee *Business Statistics* (Second Edition), Chapter XVI; McGraw-Hill Book Co., 1938. Provides a simple résumé of the analysis of time series.

Max Sasuly: *Trend Analysis of Statistics*, Chapter IX; Brookings Institution, Washington, D. C., 1934. A discussion of the smoothing of time series, for advanced students.

J. R. Stockton: *An Introduction to Business Statistics*, Chapters XI and XII; D. C. Heath and Co., Boston, 1938 Chapter XII contains a brief description of "business barometers."

CHAPTER XX

FUNDAMENTALS IN INDEX NUMBER CONSTRUCTION

Meaning and Uses of Index Numbers

Index numbers are devices for measuring differences in the magnitude of a group of related variables. These differences may have to do with the price of commodities, the physical quantity of goods produced or marketed, or such concepts as "intelligence," "beauty," or "efficiency." The comparisons may be between periods of time; between places; between like categories, such as persons, schools, or objects. Thus we may have index numbers comparing the cost of living at different times or in different countries, the physical volume of production in different years, or the efficiency of different school systems. A few uses to which index numbers are put are described below.

(1) Perhaps the most common type of index is that of the change in price level over a period of time. One use of such index numbers, with which the reader is already familiar, is that of *deflating* a value series in order to convert it into physical terms. Referring back to Chapter XIV, Table 79, we find that hourly wages were reduced to hourly real wages by dividing by an index of the cost of living. Similarly, we might wish to convert a time series representing value of construction contracts awarded to a physical basis by deflating with an index of construction costs.

(2) Price movements may be studied in order to discover their cause, or their effect on the economic community. In order to study such economic relationships, it is customary to compare changes in the price level with changes in other series, such as gold, bank reserves, bank deposits, bank debits, and the physical volume of production. Such studies may involve, not only the average change in price relatives, but also: (a) dispersion of price relatives; (b) shape of frequency distributions of price relatives; (c) alterations in the relative positions of such percentages (displacement of prices); (d) magnitude of change in price with changes in quantity offered for sale; (e) magnitude of changes in purchases or production with changes in price (elasticity of demand or supply); (f) fre-

quency with which different prices change; (g) magnitude of price changes with changes in demand.[1]

(3)   Changes in the price level may be measured in order to control them.   Thus the increase in official price of gold in 1933–1934 was in part an attempt to raise the price level.   If index numbers showed the price level to be higher after the price of gold was raised, this result might be taken as an indication that the gold policy was effective.

Occasionally, governmental influence is exercised not to raise, lower, or stabilize the price level, but to raise one group of prices relative to another. Thus the United States Government has considered various devices, and tried some, to raise agricultural prices to a "parity" with industrial prices.

(4)   Occasionally a contract is made in such a way that the effects of changes in the purchasing power of the dollar are minimized.   Thus the Philadelphia Rapid Transit Company agreed in 1926 to adjust wages annually in such a way that, regardless of price changes, the pay envelope would always support the same standard of living.   Obviously, then, it was necessary to construct an index of the cost of living of their employees in order to determine changes in wage rates.

(5)   Closely related to the use just mentioned is that of estimating for rate-making purposes the reproduction cost of utilities.   It is very laborious to re-appraise properties at frequent intervals; but, once having arrived at a satisfactory valuation as of a particular time, it is a simple matter to revise the valuation at frequent intervals in accordance with changes in the level of a price or cost of production index.   The legal status of index numbers for such purposes is in doubt; however, the United States Supreme Court has held that, if index numbers are to be so used, the particular index employed must be one that is appropriate to the particular purpose in view.   An index of "the general price level," for instance, is clearly inappropriate.[2]

(6)   Illustrations of average price comparisons among different regions are not common.   It is very difficult to make such comparisons since the relative importance of goods produced and/or consumed in the different

---

[1] Much careful study along all of these lines except elasticity of demand has been done by Frederick C. Mills.   The results have been published by the National Bureau of Economic Research in *The Behavior of Prices* (1927).   The idea of elasticity has been developed by Cournot, Alfred Marshall, and Henry L. Moore.   Henry Schultz has contributed greatly to the statistical measurement of elasticity of demand and supply See his book, *The Theory and Measurement of Demand,* the University of Chicago Press, Chicago, 1938.   Magnitude and frequency of price change have been studied statistically by Gardiner C. Means and published in *Industrial Prices and Their Relative Inflexibility,* Senate Document No. 13, 74th Congress, 1st Session.

[2] See "Index Numbers and Public Utility Valuation," by Robert W. Harbeson, *Journal of the American Statistical Association,* Vol. 31, June 1936, pp. 245–257.

places varies so widely. However, the National Industrial Conference Board has compiled an index of the cost of living in 1927 in twelve industrial cities, with the object of comparing the "differences in the cost of maintaining an established standard of living" between different regions and between different cities in the same region.

(7)   There are several organizations that compile indexes comparing physical changes over a period of time. These relate to the physical volume of trade, industrial production, factory production, sales, stocks of goods, etc. We have already used such indexes in our analysis of time series. They are extremely useful for the historical study of secular trends, seasonal variations, and business cycles, and are indispensable for persons who wish to keep abreast of current business conditions.

(8)   Forecasting indexes are compiled by most forecasting organizations. Although many of the indexes seem sound in theory, and in practice when applied to periods before they were actually used, unfortunately most of them do not work when put to current use. It is also not uncommon to find that a forecasting index works satisfactorily during periods of mild prosperity and depression but fails during a severe depression. Forecasting is discussed more fully in Chapter XXV.

(9)   It would appear from the above discussion that most indexes are price indexes. Historically they have been in use longer, and currently they are probably the most numerous. Quantity indexes are much more important than the amount of space devoted to them in paragraph 7, above, would indicate. Other varieties of indexes are diverse in nature and few in number. As an illustration of one type may be mentioned an index of school efficiency. Following the pioneer work of Leonard P. Ayres, who in 1920 published index numbers of the rating of state school systems, a number of similar studies have been undertaken.[3] Among the factors most commonly combined in the general index are: (a) school days per year; (b) per cent of school population attending schools daily; (c) ratio of high school enrollment to total enrollment; (d) average expenditure per pupil in average daily attendance; (e) average expenditure per pupil for purposes other than salaries; (f) average salary of teachers.

An index number is obtained by combining a number of variables by means of a total or an average. This statement will be clarified by reference to Table 136. In column 2 is a single price series of common building brick, and in column 3 is a series of relatives based upon these prices In column 4, however, there is a series of index numbers based on various kinds of brick and varieties of tile, which may be referred to collectively

---

[3] For an analysis and bibliography of these studies, together with a brief description of several, see "Estimating State School Efficiency," *Research Bulletin of the National Education Association*, Vol. X, No. 3, May 1932, especially pp. 104–112.

as a price *index*. These index numbers may be constructed by combining year by year, in a manner which will be described, common building brick prices and prices of other commodities in the brick and tile group. In column 3 the brick prices are expressed relative to 1926 as 100. Such a series is a series of relatives—*price relatives* in this case. Index numbers can be constructed also by averaging the price relatives of each year sep-

TABLE 136

Price and Price Relatives of Common Building Brick, and
Brick and Tile Price Index, 1926–1937

| Year | Common building brick | | Brick and tile index number (1926 = 100) |
| | Price per 1,000 | Price relative to 1926 | |
(1)	(2)	(3)	(4)
1926	$13.913	100 0	100 0
1927	14.024	100 8	95 7
1928	13.718	98 6	95 6
1929	13 621	97 9	94.3
1930	13 050	93 8	89 8
1931	12 396	89.1	83 6
1932	11 214	80 6	77 3
1933	11 047	79 4	79 2
1934	12.591	90 5	90.2
1935	12 341	88.7	89 4
1936	12 313	88.5	88 7
1937	12 647	90 9	93.5

Source: United States Bureau of Labor Statistics, *Wholesale Prices*, Bulletins
of various years. For derivation see source note of Table 138.

arately. The first method is usually referred to as the *aggregative method*, while the second is that of *averaging price relatives*. These explanations will become clearer as they are developed more fully.

Problems in the Construction of Index Numbers

Among the problems which the statistician encounters in index number construction are:

(1) Selection of series for inclusion in index.
(2) Selection of source of data.
(3) Selection of base.
(4) Method of combining data.
(5) System of weighting.

Not all of these problems are of equal importance, nor are they always independent of one another. Thus a simple system of weighting would require a different, and usually larger, list of commodities for a price index

than would a method that employs a separate weighting system for each subgroup of an index. Likewise, as will be explained later, the weighting system to use depends in part upon the method of combining the data. It is convenient to include both the method and the system of weighting in one formula, and to discuss both points in the same section. Likewise, problems 1 and 2, noted above, will be considered together. A more complete understanding of these points will result if the behavior of price relatives is considered first.

An Illustration of the Behavior of Price Relatives

The United States Bureau of Labor Statistics at the present time compiles an index of wholesale prices consisting of 813 separate commodities or series. It also computes price relatives for each of these series. From these price relatives frequency distributions have been made, and deciles subsequently computed for July of each year from 1926 to date.[4] In Chart 208 the first, third, fifth, seventh, and ninth deciles are shown. (The fifth decile is, of course, the median.)

First, it should be noticed that there is an evidence of central tendency among the movements of the price relatives, as evidenced by the fact that the central bands are generally narrower than those at the top and bottom. This suggests that the movements of prices are not entirely random, but are bound together by some underlying force of a monetary or other nature. Possibly this force is only the interdependency of prices, of which the economists speak. While this central tendency is marked in most years, it is not so clear cut in some years as in others. For instance, in 1932 the tendency is not at all apparent.

Chart 209 suggests a possible explanation of this situation. In section A are plotted indexes of farm products, foods, and all commodities other than farm products and foods. The price of farm products dropped greatly during the depression, the price of foods somewhat less, while other commodities remained fairly stable in price. Not only is agriculture composed of a large number of small scale farmers and total farm production not very responsive to price, but also the demand for farm products is inelastic. Consequently, as our export market diminished during the depression and as domestic demand fell off, it was necessary to lower the prices considerably in order to sell the crops. Section B of this chart, the indexes of

---

[4] Distributions by months, July 1927–July 1936 for all commodities have been compiled by Leonard Ascher. See "Variations in Price Relative Distributions, 1927 to 1936" by Leonard Ascher, *Journal of the American Statistical Association*, Vol. 32, No. 198, pp. 271–280. The deciles were computed by the writers from his July distributions. The deciles for building materials appearing in Chart 210 were also computed by the writers and refer to average prices for the entire year.

which overlap considerably those of section A, shows the same general picture. Although the demand for many finished products fluctuates tremendously with the business cycle, the demand at a given time is elastic, and a smaller price reduction is needed to stimulate buying. Of more importance is the fact that many manufacturing industries are composed of large scale enterprises, which are able, acting individually or collectively, to restrict the output and to maintain price.

Whatever the economic factors involved, we are forced to the conclu-
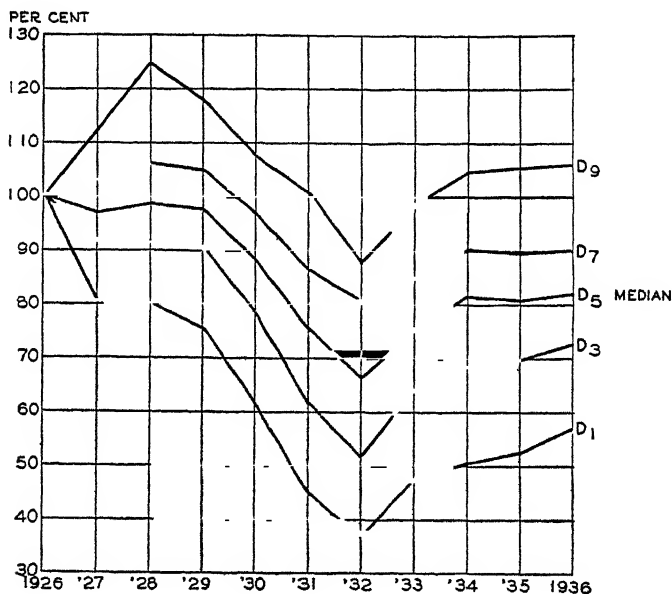
Chart 208. Dispersion of July Price Relatives as Shown by Deciles 1, 3, 5, 7, and 9, by Years, 1926-1936. (Price relatives are those for all commodities included in United States Bureau of Labor Statistics Index of Wholesale Prices. For source of data see p. 577, note 4.)

sion that a frequency distribution of price relatives is not homogeneous in character. It is a compound of separate frequency distributions, each with a characteristic mode. Under conditions of extreme economic dis location these different modes draw far apart, with the result that the compound distribution is flat topped, as is the case in 1932. As the dif ferent elements draw more nearly into the same relationship that existed in the base period, the different modes draw closer together and the central tendency becomes more marked. (This does not mean that the price system has been brought into equilibrium, a condition which would exist

if no one any longer were to make changes in his rate or method of production, his schedule of purchases, or his price bids or offers.)

A second point to observe is the dispersion. It tends to become greater as the distance from the base period increases, although this tendency is counteracted to some extent as the median value approaches that of the base period. Chart 209 shows that by 1936 the 1926 price relationships had been partially reestablished but at a lower level. Perhaps if the 1926
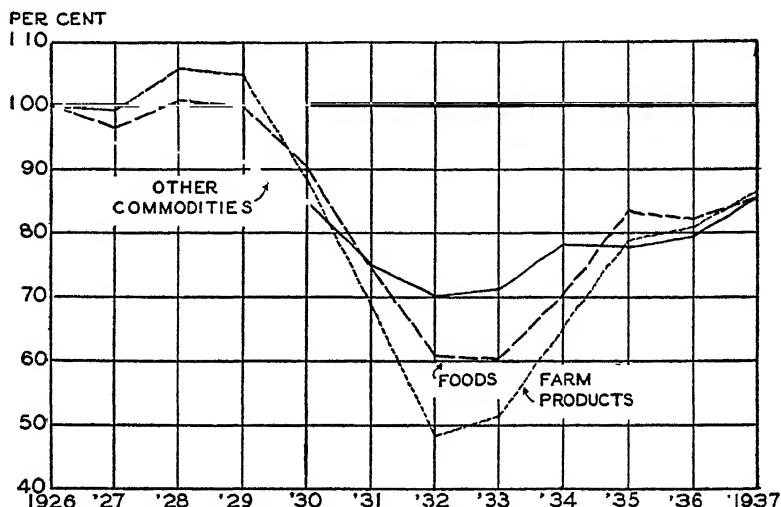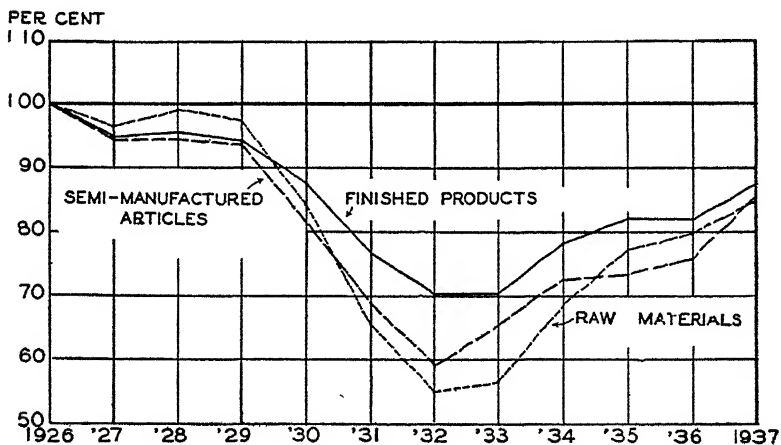
Chart 209. Major Subdivisions of United States Bureau of Labor Statistics Index of Wholesale Prices, 1926-1937. (For source of data see Table 136.)

price level were to be reached, the 1926 price relationships would again be destroyed. There is also the possibility, not clearly shown by Chart 208, that dispersion may vary with the different phases of the business cycle.

Still a third point to notice is the shape of the distributions of Chart 208. During 1927 and 1928 the skewness (judged from the relative positions of the deciles) appears to be positive. Many persons are of the opinion that this is an inherent characteristic of frequency distributions of price relatives, since they can increase indefinitely, while a selling price can decline only to zero. On the other hand, it may be suggested that price relatives are dominated more by the laws of economics than those of mathematics The limits of price advances or price declines are certainly influenced by the willingness of persons to buy at different prices. Furthermore, the direction of price change has something to do with the sign of the skewness. Beginning with 1929 the skewness begins to be negative, and the price level from this year on is definitely below that of 1926. The explanation may be that price changes are to a great extent a result of sensitive competitive price, changing to a varying extent, while managed prices tend to be sluggish.

In this chapter, building material prices have been chosen as material to illustrate index number construction, and the price quotations used are from those compiled by the United States Bureau of Labor Statistics. There are seven subgroups of building material prices: (1) brick and tile; (2) cement; (3) lumber; (4) paint and paint materials; (5) plumbing and heating; (6) structural steel; (7) other building materials.

From what has already been said, we should expect building material price relatives to form frequency distributions with characteristic central tendencies. Some of the factors which relate building prices to each other are as follows: Some building materials are complementary products, such as sand and cement; others are substitutes for each other, as brick and lumber for house exteriors; still others are joint products, as sand and gravel. In some cases a change in the price of one commodity affects another in the same direction, in other cases in an opposite direction. On the whole it seems that the forces making for uniformity of price movement are much stronger than those making for diversity. Not only should we expect a central tendency for building materials different from that of commodity prices as a whole, but we should expect less variation among the price relatives.

In general these expectations are fulfilled. Thus we find that the median price relative of the 102 building materials in 1932 is 74.6, as compared with 66.2 for all commodities (as represented by the commodities in the United States Bureau of Labor Statistics wholesale price index). Furthermore, the median of each of the building material subgroups in this year

is higher than 66.2, with the exception of lumber, which is 64.1.   Likewise, we find the spread between the first and ninth deciles to be much smaller for the building materials group and for each of its subgroups than for all commodities, except in the case of paint and paint materials.   Although we find the same tendencies in regard to skewness for building materials that we do for commodities in general, we find the distribution more peaked—that is, with a more pronounced central tendency.   Though most of the separate subgroups contain too few items to yield readily much information concerning the shape of the distribution of their price relatives,
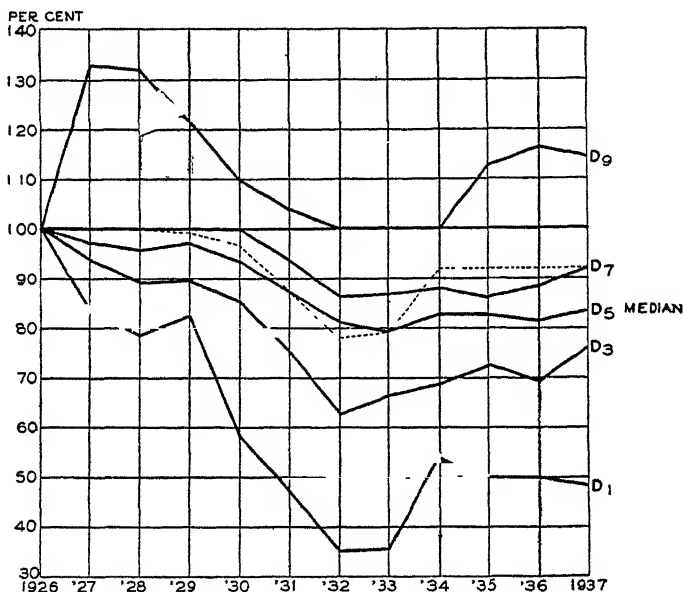
Chart 210.   Deciles of Price Relatives of Paint and Paint Materials and Price Relative of Outside Gloss White Paint (Dotted Line), 1926–1937.   (For source of data see Table 136.   Prices are average prices for each year.)

the case of paint and paint materials is interesting.   As may be seen by inspection of Chart 210, the range between the fifth and seventh deciles is very narrow, indicating a high degree of uniformity in movement for the most closely bunched fifth of the commodities.   It is interesting to note also that until 1931 the seventh decile remained at 100.   In fact, there were four of the 29 commodities which did not change in price until 1931, and one (bone black) that is still at its 1926 level.   Although the prices of a considerable proportion of tne commodities seem to be rather rigid, there are many which are very sensitive in their fluctuations.   This is indicated on the chart by the wide variation from the median of the

first and ninth deciles. As a striking illustration we find shellac, which rose to 138.6 per cent in 1927 and fell to 30.2 per cent in 1932. As in the case of the all-commodity distribution, we find the paint and paint materials distribution to be negatively skewed during the depression years.

Data for Index Numbers

Although the method of combining the variables is of considerable importance in constructing index numbers, it is insignificant when compared with the problem of selecting the data that are the raw materials of the index. Too much emphasis cannot be put upon this point. The data must be accurate and comparable, and the sample representative. A sample cannot be expected to be representative unless an adequate number of items is included. To state the idea in other language: a sufficiently large sample of relevant items must be selected to obtain reliable index numbers.

Needless to say, the commodities to be chosen for a price index, and the type of quotation to be selected, depend on what is being measured. A wholesale price index requires wholesale prices. A cost-of-living index necessitates not only retail prices of food, but rents, gas and electric rates, etc., applying to the class of persons for whom the cost of living is to be ascertained. An index of the changing cost of constructing frame houses in Atlanta, Georgia, should include those materials and items of labor that are used in frame houses built in Atlanta, and the prices should be the Atlanta prices of those materials, or the wages in Atlanta of the kind of labor used.

When selecting the sources of data for index numbers, we may rely on regularly published quotations or obtain periodic special reports from the merchants, producers, exporters, or others who possess the basic information needed. Under either circumstance we must make sure that the data pertain strictly to the thing being measured. Thus, if retail price changes are being measured, a quotation might be from a market place, an independent store, a department store, a mail order retail store, a manufacturer's outlet, etc. These different sources should not be mixed indiscriminately. Neither should first of the month quotations, middle of the month quotations, and end of the month quotations ordinarily be combined in one index.

The discussion immediately following is in part an application of principles discussed in earlier chapters of this book, especially Chapters II and XII. The great importance of the proper choice of data for index numbers justifies a bringing together of these principles, even though slight duplication is involved.

Accuracy. Some of the statistical data that appear in precise printed form cannot be depended upon. If the person or company reporting the data uses the data, they are likely to be accurate; but if the data are merely statistical reports furnished to an outside agency, they may be compiled originally by careless and indifferent clerks whose sole interest is in filling the form with ink marks as quickly as possible. It therefore behooves the statistician to ascertain how the data are collected, and to select his source with discrimination.

Comparability. Standard grades of the same commodity are, of course, comparable between different dates; however, a 1908 automobile cannot be compared with a 1938 automobile. The Automobile Manufacturers Association compiles an index of the price per pound of automobiles! If the object is to compare the amount of money that must be spent in different years to purchase an automobile which provides an equal amount of "utility" each year, the above index has a decided upward bias, since there has been a gradual increase in utility per pound. Nor is it easy to see how the price of a "standard" automobile could be computed for different years, since in not more than one year could such a standard automobile ordinarily be found   The upward bias of price quotations is greatest in the case of highly processed manufactured goods; it is present also in the case even of some agricultural commodities. It is likely, therefore, that most price index numbers have an upward bias.

A similar problem arises when one article passes out of wide use and its place is taken by a different commodity serving somewhat the same use. For instance, the stagecoach of 100 years ago has been superseded by the streamlined air-conditioned train. If we should find that the fare from Washington, D. C , to Philadelphia were the same in the two periods, we should not conclude that the cost of the same service had remained the same. Think of the saving of time required to make the trip, and the added comfort of travel provided by the modern streamlined, air-conditioned train. This problem, however, is not so difficult as the gradual change in quality of the same product. For in the period between 1830 and 1930 there was a time when both the stagecoach and the train were in use; at that time, change in cost of transportation could be measured and a substitution of trains for stagecoaches could be made for subsequent comparisons. But how should we compare the price of amusements in one country, which goes in for pulque and bull fights on a large scale, with that of another country which indulges mainly in beer and Wagner operas?

Representativeness. Since index numbers are usually obtained from samples, we must try to obtain a sample that behaves like the population from which it is drawn. Probably the most satisfactory way of accomplishing this is to divide the original data into groups and subgroups

and to draw a representative sample from each of these.   Certain statistical tests (which are described in this section) may be applied to the entire sample so selected, to determine its representative character as a whole.

As previously stated, we should expect different groups of commodities affected by different economic factors to display characteristic patterns of behavior.   For example, we should expect price (and quantity) movements of foods to be different from those of building materials.   The demand for food products is inelastic, while that for building materials (classified as durable goods, the purchase of which can be postponed) is elastic. On the other hand, the supply of foods over a short period of time is dependent to a considerable extent on the weather, while the supply of building materials is subject to conscious control of the fabricators.   Likewise, we should expect the price fluctuations of any group to be distinctive and their average movements to differ from those of all commodities taken as a whole.   So also we should expect each subgroup of a particular group (such as the paint and paint materials subgroup of the building materials group) to exhibit characteristic fluctuations.   Furthermore, the statistical data presented earlier in this chapter indicate that the facts are in accordance with the theory.

For the building materials illustration running through this chapter, seven commodities have been selected, one for each of the seven subgroups of which building materials are comprised.   (Of course, selection of only one commodity for each subgroup is a great oversimplification of the problem, and is for illustrative purposes only.)   The different groups and the commodity representative for each, together with the unit of price quotation for each, is as follows:

	Subgroup	Representative	Unit	Place of quotation
1.	Brick and tile	Common building brick	1,000	Plant
2.	Cement	Portland cement	barrel	Plant
3.	Lumber	Hard maple, No 1	1,000 board feet	Chicago
4.	Paint and paint materials. ..	Outside white gloss house paint	gallon	Plant
5.	Plumbing and heating.... .	Lavatories	each	Factory
6.	Structural steel	Structural steel	100 pounds	Mill
7.	Other building materials.. .	Building gravel	ton	Plant

In selecting the representative commodity for each subgroup, the object was primarily to choose that commodity the price behavior of which was fairly representative of the subgroup from which it was selected.   Thus in Chart 210, the dotted line representing outside white gloss house paint is seen to stay rather close in most years to the narrow band which encloses the typical items in the subgroup.   It cannot be claimed that it is

feasible to select one commodity which will adequately represent a group, but if it is possible to select only a limited number of items, preference should be given to those that conform most closely to the central tendency of the group.

Having selected commodities that are, individually or collectively, fairly representative of the group from which they were selected, it remains to be seen whether proportionate representation is obtained for each group. Table 137 indicates the extent to which each subgroup is represented in

TABLE 137

RATIO OF SAMPLE VALUE TO POPULATION VALUE OF EACH SUBGROUP OF BUILDING MATERIAL COMMODITIES, 1926

Subgroup	Thousands of dollars		Per cent of total		Ratio of sample to population (per cent)
	Population	Sample	Population	Sample	
Brick and tile	380,031	103,286	8.5	14 2	27.2
Cement	260,803	260,803	5 9	35 9	100 0
Lumber . . .	1,358,705	49,104	30 5	6 8	3.6
Paint and paint materials	634,869	87,746	14 3	12 1	13 8
Plumbing and heating	281,213	21,927	6 3	3 0	7.8
Structural steel	148,868	148,868	3 3	20 5	100.0
Other building materials	1,390,395	54,386	31.2	7.5	3 9
Total . .	4,454,884	726,120	100 0	100 0	16 3

Source See Table 136

the sample as compared with the population. Extent of representation is measured by the value marketed in 1926, the base year.

It appears that the representation of brick and tile, of cement, and of structural steel should be reduced, and that of the other four increased. It would be a simple matter to select additional commodities for these four subgroups; however, there is a difficulty in reducing the importance of cement and structural steel, each of which are represented 100 per cent. Since it is ordinarily impracticable to have each sample group co-extensive with the entire population, the remedy must be found in the weighting of the commodities.

A further test of the representativeness of the sample can sometimes be applied: Do the value changes of the sample coincide with those of the population? This test should be applied not only to the whole sample, but to the various groups and subgroups into which it is divided.[5]

[5] This test is similar to Irving Fisher's "total value criterion," which states that the price index multiplied by the quantity index should equal the ratio of change of the total value of the population. See Irving Fisher, "The Total Value Criterion," *Journal of the American Statistical Association,* Vol. XXII, December 1927, pp. 419–441.

Adequacy. It has been shown that the reliability of a mean of a random sample increases with the square root of the *number* of items included. Likewise, the larger the proportion of items included, the more reliable is the mean.[6] It would appear, then, that we should ordinarily select some of the more important items first, and as many other suitable items as resources will permit. The absolute number of items to use is a question which cannot be answered in general terms.

Selection of Base

Regardless of the formula employed for weighting and combining the data, it is customary (although not necessary) to select some period of time as 100 per cent with which to compare the other index numbers. A month is unquestionably too short a period to use as base period, since any one month is likely to be unusual on account of accidental or seasonal influences. A year, however, is often used. Before the World War 1913 was the base of the United States Bureau of Labor Statistics index of wholesale prices; but more recently it has been shifted to 1926. At the time the shift was made, it was the consensus of opinion that the post-war prices would stabilize themselves at about the 1926 level. Consequently this year was looked upon as a good year to use as a basis for comparison. However, it is probable that no one year is sufficiently "normal" to be a good basis of comparison. Business and prices are always advancing or receding with the business cycle. Though not so specific, an average of several years is a better base. The period 1910 through 1914 has sometimes been used as a price base, while the 1923–1925 average is often used for quantity indexes. A useful solution is to employ the period of years that is used by some of the other indexes with which the one being constructed is likely to be employed.*

Although a particular base may be satisfactory for a number of years, that base becomes less meaningful as time passes, and it eventually becomes desirable to shift to a more recent period. The reasons are: (1) the dispersion of price relatives becomes so great that no average is reliable; (2) the pattern of consumption changes to such an extent that no aggregate of commodities can be found which includes the major expenditures common to both periods; (3) the quality of many commodities, nominally the same, progressively changes with time. An indirect basis of comparison may be had by utilizing a chain index system. This method, which is not completely satisfactory, will be explained in the following chapter.

---

[6] It should be noted that the sample used for an index number is generally a stratified sample, and that the items from each stratum are not drawn at random. Consequently, ordinary reliability formulae are not applicable.

* The statistical agencies of the United States government are now computing several indexes on a 1935–1939 base.

TABLE 138

Construction of Simple Aggregative Index Numbers of Building Material Prices, 1926–1937

(Prices and values in dollars)

Commodity	Unit*	1926 p_{26}	1927 p_{27}	1928 p_{28}	1929 p_{29}	1930 p_{30}	1931 p_{31}	1932 p_{32}	1933 p_{33}	1934 p_{34}	1935 p_{35}	1936 p_{36}	1937 p_{37}
Common building brick	1,000	13 913	14 024	13 718	13 621	13 050	12 ???	11 ???	11 047	12 591	12 ???	12 313	12 647
Portland cement	barrel	1 744	1 686	1 672	1 601	1 601	1 ??	1 ??	1 502	1 636	1 ???	1 667	1 667
Hard maple, No. 1	1,000 board feet	55 673	52 333	54 114	54 615	50 996	37 ???	28 ???	35 519	45 095	45 ???	47 322	54 058
Outside, white, flat house paint	gallon	2 208	2 208	2 208	2 193	2 140	1 ???	2 ???	1 751	2 031	2 ???	2 031	2 031
Lavatories	each	12 374	11 174	11 161	10 407	10 679	10 ???	9 ???	8 786	8 996	7 ???	8 699	9 157
Structural steel	100 pounds	1 958	1 854	1 864	1 921	1 709	1 ??	9 ??	1 625	1 776	1 ???	1 860	2 214
Building gravel	ton	941	911	906	906	865	???	7 ??	805	858	???	854	886
Aggregate value	..	88 811	84 190	85 643	85 264	81 040	66 482	54 457	61 035	72 983	72 441	74 746	82 660
Index number (per cent of 1926)	..	100 0	94 8	96 4	96 0	91 2	74 9	61 3	68 7	82 2	81 6	84 2	93 1

* For further explanation of units, see p 584.

Source: See Table 136. Prices for years 1927 to date were obtained by multiplying 1926 prices by the percentages of Table 142. This was done because the actual prices as published in the official bulletins are not always comparable on account of substitution of new commodities, grades, or quotations for old. The percentage figures, however, are adjusted for such substitutions.

TABLE 139

Construction of 1926–1937 Aggregative Index Numbers of Building Material Prices, Weighted by Quantities Marketed in 1926

(Quantities in thousands, values in thousands of dollars)

Commodity	1926 quantity q_{26}	Value of 1926 quantities marketed at prices of specified year											
		1926 $p_{26}q_{26}$	1927 $p_{27}q_{26}$	1928 $p_{28}q_{26}$	1929 $p_{29}q_{26}$	1930 $p_{30}q_{26}$	1931 $p_{31}q_{26}$	1932 $p_{32}q_{26}$	1933 $p_{33}q_{26}$	1934 $p_{34}q_{26}$	1935 $p_{35}q_{26}$	1936 $p_{36}q_{26}$	1937 $p_{37}q_{26}$
Common building brick	7,1?4	103,290	101,11?	101,?1?	101,122	96,883	92,028	83,253	82,013	93,476	91,620	91,412	93,891
Portland cement	149,?13	260,803	2?2,?29	250,0?6	239,418	239,418	207,117	201,285	224,014	244,652	248,840	249,288	249,288
Hard maple, No. 1	?8,?	19,104	?0,1?8	1?,??1	48,170	44,978	33,341	25,289	31,128	39,774	40,510	41,738	47,679
Outside, white, flat house paint	??,?	87,746	?7,7?6	?7,746	87,150	85,044	76,698	68,631	69,5?5	80,712	80,712	80,712	80,712
Lavatories	1,???	21,927	1?,777	1?,?81	18,141	18,923	18,617	16,182	15,569	13,880	13,880	15,415	16,226
Structural steel	76,031	148,869	140,9?1	1?1,?22	146,056	129,937	123,702	120,433	123,550	135,031	136,780	141,418	168,333
Building gravel	5?,?46	51,?86	5?,??3	52,363	52,363	49,994	48,317	45,196	46,526	49,589	48,722	49,358	51,207
Aggregate value		726,125	703,560	701,215	692,720	665,177	599,820	560,269	593,185	659,175	661,064	669,341	707,336
Index number (per cent of 1926)		100 0	96 9	96 6	95 4	91 6	82 6	77 2	81 7	90 8	91 0	92 2	97 4

Source: See Table 136. Figures for $p_n q_o$ values are q_{26} quantities multiplied by prices of Table 138.

Aggregative Price Index Numbers

It has already been stated that there are two methods of constructing index numbers: (1) by computing aggregate values; (2) by averaging relatives. By the first method, as will be explained in this section, the prices or quantities are made comparable, are automatically weighted by being reduced to dollar values, and then are combined into aggregate values. In the following section the method of averaging relatives will be explained. There it will be shown that usually the two methods are merely alternative methods of obtaining the same result. The aggregative method obtains the result directly, and produces a result that has a simple and clear meaning; the other method is more roundabout, and its meaning is more technical. Nevertheless, there are situations in which the aggregative method is not applicable, and recourse must then be had to the averaging of relatives.

Simple aggregates. Table 138 illustrates the construction of a simple aggregative price index. The prices of each commodity in any given year are merely added together to give the index number for that year. It is then frequently convenient to designate some year as a base, which is set equal to 100. In this illustration all of the index numbers are expressed in the final column as a percentage of the 1926 number, found by dividing each one of the numbers by the value in the base period ($88.811) and multiplying by 100.

In Chart 211A are shown the seven components of this index and, at the top, the index itself. The vertical distance between the different curves gives some idea of their relative importance.[7] This chart shows in striking fashion why the index follows rather closely that of hard maple. That commodity was priced in 1926 four times as high per unit as the product with the next highest unit price, building brick, and comprised

$$\frac{\$55.673}{\$88.811} = 63\tfrac{1}{2}$$ per cent of the 1926 aggregate. This is far in excess of

its actual importance as a commodity used in building. It is apparent that the influence which a commodity exerts on a simple aggregative index depends on the price per unit of quotation. In this instance, hard maple was the predominant item; if house paint were quoted at wholesale by the barrel instead of by the gallon, that commodity would largely have determined the course of the index. The weighting of an aggregative index by one commercial unit of each commodity represented, then, is illogical in that it neglects to consider the actual importance of the different com-

---

[7] This chart was drawn on ratio paper in order that the proportionate changes of the different series might be compared with each other. Actually, the chart fails to give full effect to the relative influence of the higher priced series, since the vertical space between the curves is in proportion to the logarithms of the prices rather than the actual prices.
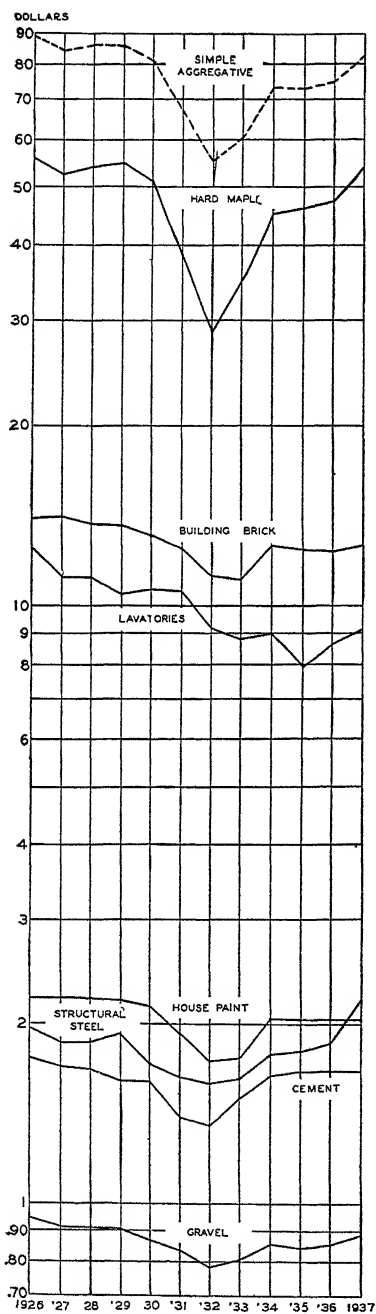
Chart 211A. Unit Prices of Seven Building Materials and Simple Aggregative Index Numbers, 1926-1937. (Data of Table 138.)
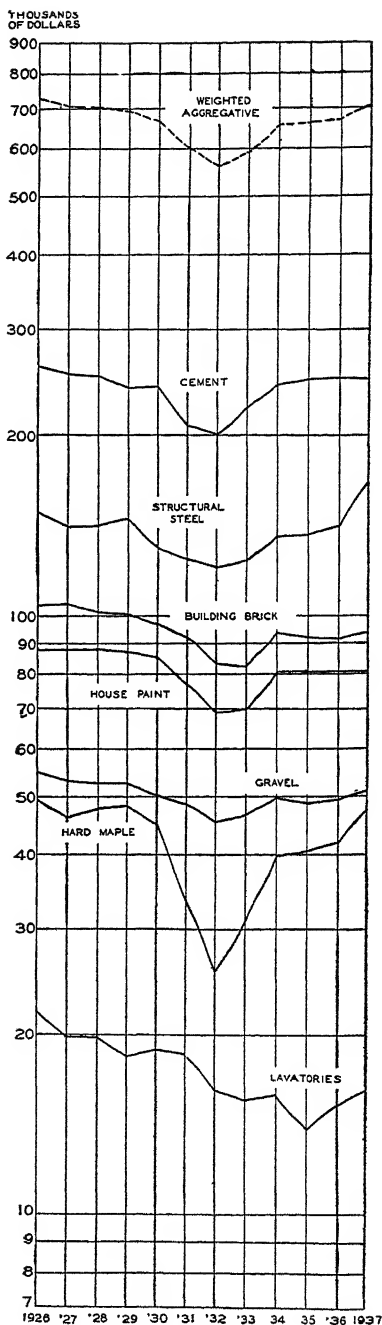
Chart 211B. Values of Base Year Quantities of Seven Building Materials and Aggregative Index Numbers with Base Year Quantity Weights. 1926-1937. (Data of Table 139.)

589

modities; it is haphazard in that the relative influence of the different commodities is determined by factors quite irrelevant to the purpose of the price index. The problem would in no sense be solved if all commodities were reduced to a price per pound, for some commodities, such as diamonds, are very costly per pound and yet are not very important in our economic life, while coal, which is of tremendous importance, is relatively cheap per pound. Furthermore, some goods, such as electric power or human labor, cannot be reduced to a pound basis. Still another solution is to take as the unit of quotation the amount that can be purchased for one dollar in the base year. But this is scarcely more logical, since it would be very unusual if the same amount of money were spent on each commodity in every year.

Before consideration of the construction of weighted aggregative index numbers, it may be helpful to state symbolically the method we have just used. The formula is

$$P = \frac{\Sigma p_n}{\Sigma p_o},$$

where P means price index.

p refers to price of an individual commodity.

o refers to the base period, from which price changes are measured.

n refers to the given period, the year being compared with the base. Now if the formula for a particular year (say 1931) is to be stated, it could be written

$$P_{31} = \frac{\Sigma p_{31}}{\Sigma p_{26}}.$$

These are the notations used in Table 138.

Weighted aggregates. In order to allow each commodity to have a reasonable influence on the index, it is advisable to use a weighted rather than a simple (unweighted) aggregate of prices. To construct a weighted aggregative index, a list of definite quantities of specified commodities is taken, and calculations are made to determine what this aggregate of goods is worth each year at current prices. Obviously the process is merely that of multiplying each unit price by the number of units and summing the resulting values for each period. The procedure, using the quantities marketed in 1926 as multipliers, is illustrated in Table 139. The reader, having followed the reasoning to this point, will realize now that *aggregative index numbers of price measure the changing value of a fixed aggregate of goods.* Since the total cost or value changes while the components of the aggregate do not, these changes must be due to price changes. It appears that this type of index number measures the very thing sought if we wish to determine changes in the cost of living. The Philadelphia Rapid

Transit Company, therefore, chose this type as a guide to its wage policy. The general formula for the aggregative price index is

$$P = \frac{\Sigma p_n q}{\Sigma p_o q}.$$

The symbols are those used earlier, but a new one has been added: q refers to the quantity of the commodity produced, marketed, or consumed (that is, the quantity weight, or multiplier). Since the index numbers constructed in Table 139 were weighted by base year quantities, we may write the formula more specifically

$$P = \frac{\Sigma p_n q_o}{\Sigma p_o q_o}.$$

If the reader will turn to Chart 211B, he will be struck by the wide variation between the simple and the weighted index numbers. Furthermore, the reason for the difference will at once be apparent. In the simple aggregative type, hard maple, the commodity that declined most greatly, is of dominating importance; whereas, when weights are introduced, cement becomes most important, although no longer does any single commodity exercise overwhelming influence on the course of the index. Nevertheless, it is significant that the two most important commodities, cement and structural steel, are the two commodities that have most nearly regained their 1926 level.

Selection of weights. Although in the preceding illustration the quantities marketed in 1926 were used as weights, this simple procedure is but one of several possible systems. It would have been just as easy to have taken, say, 1927 quantities as weights. If the quantity of each commodity marketed changed from year to year in the same proportion, it will make no difference to what period the weights refer, for the results will be identical. In fact, however, the relative importance of the different commodities is constantly changing, and this is due in part to the change in the relative prices of the different commodities. Therein lies a great source of difficulty for which there is no completely satisfactory solution. The answer depends in part on what the analyst thinks a price index is supposed to do.

One view is that such an index number measures the changing cost of a constant aggregate of goods. Another view concerns itself not with the goods level of analysis but with the satisfactions level; an index number should measure the changing cost of aggregates of goods yielding the same utility or satisfaction at two periods, or two places. Thus, suppose we compare the cost of living of two groups of similar persons at two periods (or places), these groups having at the two periods (or places) the same

tastes and capacity for enjoyment, as well as an income that will purchase, and does purchase, the same amount of satisfaction.[8] The commodities, of course, will be different, but if the expenditures were $2,000 the first year and $2,400 the second year, we may conclude that the cost of living has gone up 20 per cent. It goes without saying that no one has accurately made a measurement of this kind. Although it seems feasible to measure only the varying value of a fixed aggregate of goods, yet the analyst should select a list of goods that will avoid the certainty of bias in a known direction with respect to the cost of obtaining equal satisfactions at different times. The following suggestions have been made for solving this knotty problem.

1. *Use base period quantities as weights.* This is the method we have used for illustrative purposes in Table 139. However, even if there has been no change in the tastes or environment of purchasers between the two periods, purchases of those commodities that have increased relatively in price will decline relatively, and purchases of commodities that have decreased relatively in price will increase relatively. It is entirely possible that this type of index might record an increase in the price level, whereas by increasing the relative amounts purchased of commodities that decline in price, the same amount of satisfaction might actually be bought by a given individual at a lower total cost. This type of index, then, has in a sense an upward bias. It might be said that this index marks an upper limit to the price change. This method is sometimes known as *Laspeyres' method* and, as previously stated, can be defined symbolically,

$$P = \frac{\Sigma p_n q_o}{\Sigma p_o q_o}.$$

2. *Use given period quantities.* That is, use the weights that pertain to the year the price level of which is to be compared with that of the base period. This method involves the selection of a new set of weights each year, or even more often. But frequently it is impossible to obtain current quantity weights, and, even if they are available, the labor of computation is approximately doubled. Furthermore, although each period is thereby directly comparable with the base year, the comparison of the different years among themselves is not valid, for the reason that the aggregate of goods differs each year.

If we think of 1926 as being the base period, the base year weighting system answers the question: If it cost me $100 a month to live in 1926, how much would it cost me this year to live the way I did that year? The given year weighting system answers a different question: If I could have

[8] See J. M. Keynes, *A Treatise on Money*, Vol. I, pp. 96–99. Harcourt, Brace, & Co., New York, 1930.

supported my *present* scale of living in 1926 with $100 per month, how much must I spend this year?   A theoretical objection to asking such a question is that undue weight is given to the commodities that have declined in price.   It is the relative decline in price that may be responsible for their increased purchase, and, although it is price change which we are trying to measure, yet our weighting is partly determined by relative price changes.   Thus this method may be said to have a downward bias, and marks the lower limit of price change.   It is sometimes known as *Paasche's method* and has the following formula:

$$P = \frac{\Sigma p_n q_n}{\Sigma p_o q_n}.$$

3.   *Use the average (or total) quantities of base and given years.*   This is a compromise solution, although it is one which has no general bias in any known direction.   But again, as in method 2, we have shifting weights and a resulting lack of comparability among the different years.   The method was proposed independently by the English economists Marshall and Edgeworth, and the formula

$$P = \frac{\Sigma p_n (q_o + q_n)}{\Sigma p_o (q_o + q_n)}$$

is sometimes called the *Marshall-Edgeworth formula.*

4.   *Average together the quantities for all the years which the index numbers include.*   Though perhaps an excellent solution for a historical study, this plan is impracticable if the index is to be kept up to date, since it means current revision of weights and continuous recomputation of the complete set of index numbers.

5.   *Average together the quantities of several years which are thought to be typical.*   This again is a compromise solution, but it is practical and is very frequently adopted.   The list of quantities used will, however, eventually become obsolete.   When that is the case, a new index can be constructed and spliced to the old one.   Methods for so doing will be considered in the following chapter.   The construction of an index number of 1931 building material prices, using as weights the average quantity marketed in 1927 and 1929, is illustrated in Table 140.   The index number varies only one-tenth of a point from that employing base year weights. The formula for this particular index number may be written

$$P = \frac{\Sigma p_n q_{27,29}}{\Sigma p_o q_{27,29}}.$$

6.   *Determine the highest common factor.*   The weights are the quantities of each commodity common to each year, either to the base and given year, or to all the years under comparison.   In the latter case this would

mean that, for any commodity, the smallest amount marketed in any of the years under comparison would be taken. Usually, then, the quantities of the different commodities taken would not each be for the same year. This ingenious device has been suggested by J. M. Keynes[9] to avoid the sort of bias inherent in methods 1 and 2, already described. Its virtue is its modesty: the device avoids trying that which cannot be done perfectly. However, if the values of quantities that are common to the dif-

TABLE 140

CONSTRUCTION OF 1931 WEIGHTED AGGREGATIVE INDEX NUMBER OF BUILDING MATERIAL PRICES, USING 1927 AND 1929 AVERAGE QUANTITY WEIGHTS

(Quantities and values in thousands)

Commodity	Average quantity marketed in 1927 and 1929 $q_{27,29}$	1926 price per unit p_{26}	1931 price per unit p_{31}	1927, 1929 quantities at:	
				1926 prices $p_{26}q_{27,29}$	1931 prices $p_{31}q_{27,29}$
Common building brick	6,348	$13.913	$12.396	$ 88,320	$ 78,690
Portland cement .	171,926	1 744	1 385	299,839	238,118
Hard maple, No 1 .	794	55.673	37 802	44,204	30,015
Outside white gloss house paint.	49,082	2 208	1 930	108,373	94,728
Lavatories	1,590	12 374	10 506	19,675	16,705
Structural steel....	90,970	1.958	1 627	178,119	148,008
Building gravel	80,666	.941	836	75,907	67,437
Aggregate value	...	.	. .	$814,437	$673,701
Index number	..	..	. .	100 0	82 7

Source: See Table 136

ferent periods are small compared with total expenditures, or if they constitute in different periods a varying per cent of the total, or if the satisfaction derived from this aggregate of goods varies, the method is no more accurate and, quite likely, is less accurate than method 5.

7. *Make two index numbers, each with a different set of weights, and average the two together, usually geometrically.* The two systems of weighting chosen are ordinarily base and given year weights. The formula then becomes

$$P = \sqrt{\frac{\Sigma p_n q_o}{\Sigma p_o q_o} \times \frac{\Sigma p_n q_n}{\Sigma p_o q_n}}.$$

It is frequently called Fisher's "ideal" index number, because it conforms to certain tests of consistent behavior which Irving Fisher considers ap-

[9] *Ibid.*, pp. 105–109.

propriate.[10]   On the other hand, it is difficult to say precisely just what such an index number does measure.

A general criticism of any weighting system which involves the use of a different set of weights for each index number is that, although each index number may validly be compared with that of the base year, logically the index numbers of no other two years (such as 1936 and 1937) can be compared with each other.   This criticism applies to given year weights, to the average of base and given year weights, to the highest common factor method when the quantities selected are common only to the two years being compared, and to the "ideal" index number.   It does not apply to base year weights, average weights of all years, typical weights, or the highest common factor method when the quantities common to all years are used.

Although the theory of weight selection is interesting and involves logical analysis of a high order, it is easy to overestimate its practical importance. Consider the following results obtained from the building material data:

	1931 index number
System of weighting	
Simple..	74 9
1926 quantity weights (base year weights)	82 6
1927 and 1929 average quantity weights	82 7
1931 quantity weights (given year weights)	82.5
"Ideal" index number	82.6

In this case there is a very great difference between the simple and the weighted index numbers, but practically no difference between the systems of weighting.   If, however, both the prices and quantities had varied greatly in their relative magnitude, the different weightings might have given markedly different results.   Furthermore, it is usually of slight importance whether exact weights are used, or only approximate weights. Thus, Table 141 is exactly like Table 140 except that the quantity weights are rounded to one digit, but the results vary by only two-tenths of a point.   For all practical purposes, sufficiently accurate results will usually be obtained if exact weights are given to the few more important commodities, and rounded weights to the numerous unimportant commodities.[11]

Although only approximate accuracy is necessary in choosing weights, accuracy in price quotations is, in practice, of much greater importance.

---

[10] See Irving Fisher, *The Making of Index Numbers*, p. 220, Houghton Mifflin Company, Boston, 1927.   In Chapter IV Professor Fisher discusses these tests.

[11] Irving Fisher recommends that the quantities be rounded to 1, 10, 100, or 1,000. This, of course, materially lightens the work.   In rounding any quantity between 1 and 10 (for instance), the dividing point is not the arithmetic mean of these two numbers, but the geometric mean, 3.1623, since this involves the smallest *relative* error See *ibid.*, pp. 346 and 432.

If all prices moved in the same direction and at the same rate, it would make no difference what system of weighting were chosen. We have found that, in fact, distributions of price relatives do display a central tendency. But if it so happens that commodities which are changing *greatly* in relative importance during the period are also undergoing price changes materially different from the average, then the matter of weighting becomes important.

Over a number of years various changes take place: commodities shift considerably in their relative importance; old commodities disappear from

TABLE 141

CONSTRUCTION OF 1931 AGGREGATIVE INDEX NUMBER OF BUILDING MATERIAL PRICES,
WEIGHTED BY 1927 AND 1929 AVERAGE QUANTITIES ROUNDED TO ONE DIGIT

(Quantities and values in thousands)

Commodity	Average quantity marketed in 1927 and 1929 $q_{27,29}$	1926 price per unit p_{26}	1931 price per unit p_{31}	1927, 1929 quantities at:	
				1926 prices $p_{26}q_{27,29}$	1931 prices $p_{31}q_{27,29}$
Common building brick	6,000	$13 913	$12 396	$ 83,478	$ 74,376
Portland cement . . .	200,000	1 744	1 385	348,800	277,000
Hard maple, No 1 . .	800	55 673	37 802	44,538	30,242
Outside white gloss house paint	50,000	2 208	1 930	110,400	96,500
Lavatories.	2,000	12 374	10 506	24,748	21,012
Structural steel	90,000	1 958	1 627	176,220	146,430
Building gravel.	80,000	941	836	75,280	66,880
Aggregate value . . .	.	. .		$863,464	$712,440
Index number	. .			100 0	82 5

Source: Table 140

use and are succeeded by new commodities; models, styles, or grades of a commodity become obsolete and cease to be manufactured, with new models, styles, or grades taking their place; marketing centers shift, so that a price quotation at the new center must replace that at the old; f.o.b. price quotations may give way to delivered prices. Under any of these circumstances it may be desirable to express each index number, not as a percentage of the original base, but as a percentage of the preceding period. Such a link relative index number might employ any of the formulae given above, utilizing weights pertaining to either or both of the years or months being compared. Frequently these separate percentages (link relative index numbers) are chained back to the original base by a process of successive multiplication. Such an index, known as a *chain*

index, will be further described in the following chapter. Overlapping price data are needed for only a single period, as a direct comparison is made only between the prices of the current period and those of the preceding period.

Averages of Price Relatives

A brief illustration will indicate the method of obtaining index numbers by averaging price relatives.

1. *Reduce the actual prices to a percentage of the base period.* The per-
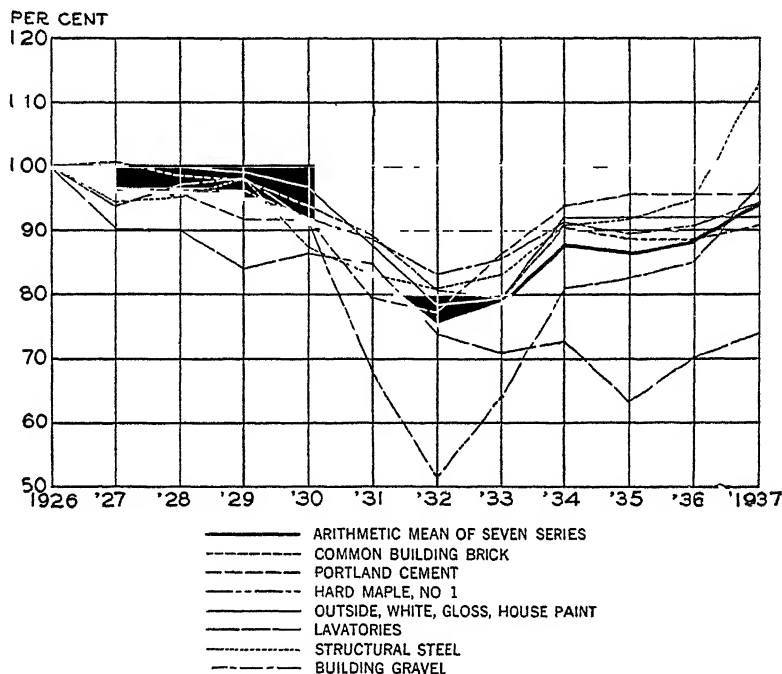
PER CENT

Chart 212. Price Relatives and Simple Arithmetic Average Index Numbers of Seven Building Materials, 1926-1937. (Data of Table 142.)

centages are called price relatives, since they are expressed not as dollars and cents but as percentages relative to the price during a certain period. Table 142 shows the price relatives for our seven building materials from 1926 through 1937. Each of these series of relatives was computed in the same manner as were the relatives for common building brick in Table 130.

2. *Average the price relatives for each year separately,* thus obtaining a series of index numbers. In Table 142 a simple arithmetic mean is used. Chart 212 shows the movement of the individual price relatives, together with the average movement. It is, of course, possible to use other types

TABLE 142

Construction of Index Numbers of Building Material Prices by Simple Arithmetic Means of Price Relatives, 1926–1937

Commodity	1926	1927	1928	1929	1930	1931	1932	1933	1934	1935	1936	1937
Common building brick	100.0	100.8	98.6	97.9	93.8	89.1	80.6	79.4	90.5	88.7	88.5	90.9
Portland cement	100.0	96.7	95.9	91.8	91.8	79.4	77.2	86.1	93.8	95.4	95.6	95.6
Hard maple, No. 1	100.0	94.0	97.2	98.1	91.6	67.4	51.5	63.8	81.0	82.5	85.0	97.1
Outside white gloss house paint	100.0	100.0	100.0	99.3	96.9	87.4	78.2	79.3	92.0	92.0	92.0	92.0
Lavatories	100.0	90.3	90.2	84.1	86.3	84.9	73.8	71.0	72.7	63.3	70.3	74.0
Structural steel	100.0	94.7	95.2	98.1	87.3	83.1	80.9	83.0	90.7	91.9	95.0	113.1
Building gravel	100.0	96.8	94.3	96.3	91.9	88.8	83.1	85.5	91.2	89.6	90.8	94.2
Average	100.0	96.2	96.2	95.1	91.4	82.9	75.0	78.3	87.4	86.2	88.2	93.8

Source. See Table 136.

TABLE 143

Construction of Index Numbers of Building Material Prices by Arithmetic Means of Price Relatives Weighted by Base Year Values, 1926–1937

(Values in thousands of dollars)

Commodity	1926 value $p_{26}q_{26}$	1926 value multiplied by price relative of specified year											
		1926 $\frac{p_{26}}{p_{26}}q_{26}$	1927 $\frac{p_{27}}{p_{26}}q_{26}$	1928 $\frac{p_{28}}{p_{26}}q_{26}$	1929 $\frac{p_{29}}{p_{26}}q_{26}$	1930 $\frac{p_{30}}{p_{26}}q_{26}$	1931 $\frac{p_{31}}{p_{26}}q_{26}$	1932 $\frac{p_{32}}{p_{26}}q_{26}$	1933 $\frac{p_{33}}{p_{26}}q_{26}$	1934 $\frac{p_{34}}{p_{26}}q_{26}$	1935 $\frac{p_{35}}{p_{26}}q_{26}$	1936 $\frac{p_{36}}{p_{26}}q_{26}$	1937 $\frac{p_{37}}{p_{26}}q_{26}$
Common building brick	103,286	103,286	104,112	101,840	101,117	96,882	92,028	83,249	82,009	93,474	91,615	91,408	93,887
Portland cement	260,803	260,803	252,197	250,110	239,417	239,417	207,078	201,340	224,551	244,633	248,806	249,328	249,328
Hard maple, No. 1	49,104	49,104	46,158	47,729	48,171	44,979	33,342	25,289	31,328	39,774	40,511	41,738	47,680
Outside white gloss house paint	87,746	87,746	87,746	87,746	87,132	85,026	76,690	68,617	69,583	80,726	80,726	80,726	80,726
Lavatories	21,927	21,927	19,800	19,778	18,441	18,923	18,616	16,182	15,568	15,941	13,880	15,415	16,226
Structural steel	148,868	148,868	140,978	141,722	146,040	129,962	123,709	120,434	123,560	135,023	136,810	141,425	108,370
Building gravel	54,386	54,386	52,046	52,374	52,374	49,981	48,295	45,195	46,500	49,600	48,730	49,382	51,232
Aggregate value	726,120	726,120	703,637	701,299	692,692	665,170	599,758	560,306	593,099	659,171	661,078	669,422	707,449
Index number (per cent of 1926)	...	100.0	96.9	96.6	95.4	91.6	82.6	77.2	81.7	90.8	91.0	92.2	97.4

Source: See Table 136. Figures for 1927–1937 are derived from $p_{26}q_{26}$ column and relatives of Table 142

of averages, such as the harmonic mean, the geometric mean, or the median. If a weighted average is used, the weights are *value* weights, as contrasted with the aggregative method, which makes use of *quantity* weights.

Type of average. A fairly good theoretical argument can be built up for the use of the geometric mean in averaging price relatives. Let us assume the simple case of measuring the difference in price level between two countries, in which two commodities only are used.

Commodity	Country A		Country B	
	Unit price	Price relative (per cent)	Unit price	Price relative (per cent)
Wheat	$0.80	100	$1 60	200
Cotton . .	12	100	06	50
Average		100		125

According to this method of calculation, the price level is 25 per cent higher in Country B than in Country A. But the reader can easily verify that, if Country B had been taken as the base and the price in Country A calculated relative to Country B, the price level in Country A would have appeared 25 per cent higher than in Country B. An unweighted arithmetic mean is therefore sometimes said to have an upward bias. On the other hand, the geometric mean of 2.00 and .50 is 1; hence the results are consistent no matter which country is considered the base.

This paradox is due to a concealed change in the weighting system. Actually there is no such thing as an unweighted index; the weights are there, be they appropriate or otherwise. Now in the table above the weight may be thought of as $1.00 for each commodity consumed, and which represents, for Country A, the base:

$1\frac{1}{4}$ bushels of wheat @ $0.80 = $1.00.
$8\frac{1}{3}$ pounds of cotton @   .12 =   1.00.

Keeping the same quantity element in our weights, we may compute new weights to be used when Country B is taken as 100:

$1\frac{1}{4}$ bushels of wheat @ $1.60 = $2.00.
$8\frac{1}{3}$ pounds of cotton @   .06 =   .50.

Let us now compute a weighted index number with Country B = 100.

Commodity	Value weight	Price relative		Weighted relative	
		B	A	B	A
Wheat	$2 0C	100	50	200	100
Cotton	.50	100	200	50	100
Total.	$2.50	...	..	250	200
Index number.. .	...	...	.	100	80

The results are now seen to be consistent. Retaining the quantity elements in the weights constant in this manner, the price level in Country B is 125 per cent of A, and in Country A it is 80 per cent of B. The so-called bias of the arithmetic mean turns out to be a matter of improper weighting. The arithmetic mean argues that, if we purchase the same commodities in the same relative quantities in the two countries, the index number is 25 per cent higher in Country B than in Country A; the geometric mean, to yield consistent results, requires that the value of the different commodities purchased be in the same ratio in the two countries (thus necessitating that in Country B a relatively smaller quantity of wheat be purchased and a relatively larger quantity of cotton).

A closely related argument for the geometric mean which is sometimes advanced is based upon the assertion that frequency distributions of price relatives tend to form a normal distribution when plotted on paper having a logarithmic X scale (or when the logarithms of the price relatives are plotted on arithmetic paper). The reasoning runs as follows: the doubling of a price represents as important a divergence (and is as likely to occur) as a decline to one-half of its former level; it is as likely to increase to $\frac{3}{2}$ of the base period as to fall to $\frac{2}{3}$ of the base period; it is as likely to rise to infinity as it is to fall to zero. The resulting frequency distribution therefore tends to be normal geometrically, and the geometric mean, which coincides with the mode of such a distribution, is the appropriate average. This argument is logical but is based upon premises that are not fully established. We are not sure that a price is as likely to double as to drop one-half. It is frequently easier to cut off buyers by doubling prices than to attract more purchases by cutting prices one-half. In the early part of this chapter the deciles of a number of price distributions were shown. In nearly every year the distribution was skewed negatively. It is noteworthy also that in every year the price level was below that of 1926. In fact, there seemed to be a rough tendency for these distributions of relatives to be skewed in the direction of the change from the base. This is not surprising in view of the much publicized tendency

for many prices to be rigid.   The change in the price level is perhaps ac-
counted for in large part by the price movements of the flexible prices.
If it be true that rigid prices resist downward changes more persistently
than they do upward, we might expect logarithms of price relatives to be
skewed negatively on the average.   At any rate, most distributions are
not normal geometrically, regardless of what may be true on the average.

It should not be thought that the geometric mean must never be used;
it merely is to be doubted that it has any inherent general superiority
over the arithmetic mean.   It is the belief of the authors that the average
to use is determined in large part by the use for which the index numbers
are intended.   If, as is very often the case, we wish to compare the amount
of money required at two different times or in two different places to pur-
chase the same commodities (or perhaps the same amount of satisfaction
by like individuals, with tastes and environment held constant), the
weighted arithmetic mean should be used.   This is because (as will be
shown) such an index number may also be regarded as a weighted aggrega-
tive index number.   On the other hand, if the primary object is the study
of price relatives, including their average behavior, the geometric mean
may be useful.

The mode is seldom advocated.   If the mode of price relatives of paint
and paint materials had been calculated, it would have remained at 100
from 1926 perhaps through 1930!   But after a number of years have
elapsed, and when an index covering a broad field is sought, it is likely
that the central tendency will not be sufficiently marked to justify use of
the mode.   The median is seldom used either, but might be appropriate
if the accuracy or representative character of some of the data is in doubt.
The harmonic mean has been suggested by Ferger (see footnote 2, Chapter
XXI) if it is desired to use the reciprocal of the price index as an index
of the purchasing power of money.

Weighting systems.   In the illustration in Table 143, the values mark-
eted in the base year (1926) are used as the weights.   Like any weighted
average, this one is obtained by: first, multiplying the relatives by their
weights; second, summing these figures year by year; and finally, dividing
these totals for each year by the sum of the weights.   The results are the
same as those obtained for the aggregative index with base year quantity
weights.   The reason is obvious.   Take a single commodity, building
gravel:

Value of 57,796,000 tons @ $0.941 (1926 price) = $54,386,000;
Value of 57,796,000 tons @ $0.911 (1927 price) = $52,652,000.
Price relative ($.911 ÷ $.941) = .96812, or 96.812 per cent;
$54,386,000 × .96812 (the value in the base year) = $52,652,000.

(Table 143 shows \$52,646,000 instead of \$52,652,000 for 1927 because the 1927 relative was taken as 96.8.)

This relationship is true, not only for each individual commodity, but for the aggregate values.[12] In symbols:

$$\frac{\dfrac{p_n}{p_o} p_o q_o}{p_o q_o} = \frac{p_n q_o}{p_o q_o},$$

$$\frac{\Sigma \dfrac{p_n}{p_o} p_o q_o}{\Sigma p_o q_o} = \frac{\Sigma p_n q_o}{\Sigma p_o q_o}.$$

Evidently the method of weighted average of relatives is usually a roundabout method of doing what may more easily be accomplished by direct means using aggregates. Furthermore, the meaning of an aggregative index seems clearer to most persons than does an average of relatives. Why, then, should not the aggregative method always be used? One reason is that the price relatives themselves are occasionally worth

[12] More generally, the following relationships may be stated with regard to price index numbers:

(1) An arithmetic average of relatives weighted by base year values $(p_o q_o)$ is the equivalent of an aggregative index weighted with base year quantities.

(2) Similarly, an arithmetic average of relatives weighted by the product of base year prices and given year quantities $(p_o q_n)$ is the equivalent of an aggregative index weighted with given year quantities.

(3) A harmonic average of relatives weighted by given year values $(p_n q_n)$ is the equivalent of an aggregative index weighted with given year quantities. Thus

$$1 \div \frac{\Sigma \left(\dfrac{1}{p_n \div p_o} p_n q_n \right)}{\Sigma p_n q_n} = 1 \div \frac{\Sigma \left(\dfrac{p_o}{p_n} p_n q_n \right)}{\Sigma p_n q_n}$$

$$= \frac{\Sigma p_n q_n}{\Sigma \left(\dfrac{p_o}{p_n} p_n q_n \right)} = \frac{\Sigma p_n q_n}{\Sigma p_o q_n}.$$

(4) Similarly it may be shown that a harmonic average of relatives weighted by the product of base year quantities and given year prices $(p_n q_o)$ is the equivalent of an aggregative index weighted with base year quantities.

These generalizations may be stated in the form of guides to the construction of index numbers, when the index numbers are to be constructed from relatives:

(a) If it is desired to use the arithmetic average of relatives, the value weights should be the products of the base prices and whatever quantities are desired.

(b) If it is desired to use an average of relatives employing value weights that are the product of given year prices and quantities of some period, the harmonic average should be used.

Under no circumstances should the arithmetic average of relatives be used with values involving given year prices, since this gives extra weight to a commodity merely because it has gone up in price. Such a procedure results in an upward bias.

studying, not only because an individual series may hold special significance for the reader, but because a study of groups of relatives may assist in selecting a sample or determining what group indexes to make.   In connection with frequency distributions it was observed that an average never gives a complete picture of any situation.   Other measures may be worth making.   Another reason is that the series to be combined can sometimes be obtained only in the form of relatives; for instance, Snyder's Index of the General Price Level (published by the Federal Reserve Bank of New York) is a weighted arithmetic average of a number of component price indexes.   The component indexes are: retail food prices; rents; other cost-of-living items; prices of industrial commodities at wholesale; farm prices at the farm; transportation costs; realty values; security prices; equipment and machinery prices; hardware prices; automobile prices; composite wages.   The use of relatives is more common, however, in constructing various types of quantity indexes, since the components of these indexes are often themselves index numbers or other types of relatives.

Commodity weights versus group weights.   The same practical advice may be offered concerning value weights that was given concerning quantity weights—only approximate accuracy is necessary.   Nevertheless, the following consideration becomes important when only a limited number of commodities is chosen: Should the value weight selected for any given commodity be the value of *that commodity* entering the market, or should it refer to the whole *group* of commodities which the commodity represents? This is likely to be a far more important consideration, over relatively short periods of time, than the question of the period to which the weights refer. The answer to this question is that, unless it is practicable to increase the number of items in some groups (and perhaps decrease the number in others) sufficiently to obtain proportionate value representation for the different groups, it is decidedly better to adjust the weights of the different items so as to obtain such group representation.

In Table 144 we have the application of the estimated total value of the different subgroups marketed in 1926 to the individual commodity price relatives of Table 142.[13]   The primary effect is to increase the weight of hard maple, the price of which declined greatly during the depression. Of considerable importance is the reduced weight of cement and structural steel, the prices of which had by 1936 approximated their 1926 levels. Gravel is increased in importance.   Although this commodity did not decline in price greatly during the depression, its general trend throughout the period has been downward.   The net result has been that the

---

[13] If an aggregative index is used, weighting quantities in accordance with the importance of their group necessitates the use of derived quantity weights expressed in abstract units.   This is done, as shown in the following table, by dividing the market

index with subgroup weights declined more during the depression than did that with commodity weights, and has regained less of its loss since 1932. It should not be concluded from the preceding illustration that paucity of price quotations is desirable. Most satisfactory results will be obtained if we select as large a number of commodities from each group as feasible, and at the same time give additional weight to those elements that are under-represented.

Another method of accomplishing the same result is to select as many commodities as convenient for each group, to compute separate group indexes, and then to combine the group indexes into a general index, using the appropriate weights. Since the group indexes are relatives, their combination presents no new problem.

It might further be noticed that weighting of commodities may in a sense be regarded as a substitute for selecting the number of commodities from the different groups in proportion to the value of those groups. For instance, we might select commodities as follows from the different subgroups:

Brick and tile	2 items
Cement	1 item
Lumber.	7 items
Paint and paint materials . .	3 items
Plumbing and heating . ..	1 item
Structural steel	1 item
Other building materials	7 items
All building materials ...	22 items

value of the subgroup by the price per unit of the commodity representing the subgroup. The derived quantities so calculated are applied in the usual fashion, in the same manner as in Table 139

COMPUTATION OF DERIVED QUANTITY WEIGHTS OF SEVEN INDIVIDUAL COMMODITIES TO CORRESPOND WITH THE IMPORTANCE OF THEIR SUBGROUPS, BUILDING MATERIALS, 1926

Commodity subgroup (1)	Value marketed (thousands of dollars) (2)	Unit price of subgroup representative (dollars) (3)	Derived quantity marketed (thousands) [Col 2 ÷ Col. 3] (4)
Brick and tile....	380,031	13.913	27,315
Cement..........	260,803	1.744	149,543
Lumber	1,358,705	55.673	24,405
Paint and paint materials .. .	634,869	2.208	287,531
Plumbing and heating .. .	281,213	12 374	22,726
Structural steel	148,868	1.958	76,031
Other building materials	1,390,395	.941	1,477,572

Source: Tables 137 and 138.

If the most important items in each group are selected, an unweighted average of price relatives will yield reasonably good results. For instance, such an index for 1931 was found by the writers to be 80.8 per cent, which is not greatly different from the figure (81.3) obtained by the use of subgroup value weights.

Before leaving this section, it is interesting to compare the results of the major illustrations developed thus far with each other and with the United States Bureau of Labor Statistics index of building material prices.

PER CENT

U S BUREAU OF LABOR STATISTICS INDEX
AGGREGATIVE WITH BASE YEAR WEIGHTS
SIMPLE ARITHMETIC AVERAGE OF RELATIVES
AVERAGES WEIGHTED BY BASE YEAR SUBGROUP VALUES
SIMPLE AGGREGATIVE
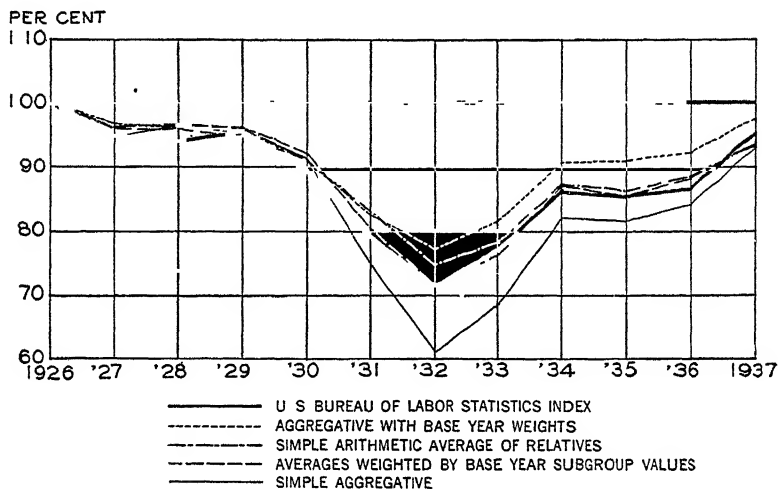
Chart 213. Index Numbers of Building Materials as Obtained by Different Methods, 1926-1937. (Data of Tables 138, 139, 142, and 144. United States Bureau of Labor Statistics Index Numbers were obtained from the Bureau's bulletin, *Wholesale Prices*, December and Year 1937, p. 9.)

The latter includes the seven items here used and a great many additional ones. Although the Bureau's weighting system is one which employs shifting weights, it affords a standard of comparison for index numbers constructed from the sample. It is easily apparent from inspection of Chart 213 that the weighting systems in the present instance rank in order of closeness of conformity as follows:

(1) Base year subgroup value weights (Table 144).
(2) Simple arithmetic average of relatives (Table 142).
(3) Base year commodity weights (Table 139 or 143).
(4) Simple aggregative (Table 138).

Although the results of this experiment are not to be taken as conclusive, they do point to the desirability of giving each important element in the system approximately its correct weight, and they indicate the unfortunate

TABLE 144

CONSTRUCTION OF INDEX NUMBERS OF BUILDING MATERIAL PRICES BY ARITHMETIC MEANS OF PRICE RELATIVES, WEIGHTED BY VALUE OF ENTIRE SUBGROUP MARKETED IN BASE YEAR, 1926–1937

(Values in thousands of dollars)

Commodity	1926 value p_0q_0	1926 subgroup value multiplied by price relative of specified year											
		1926	1927	1928	1929	1930	1931	1932	1933	1934	1935	1936	1937
Brick and tile	380,031	380,031	383,071	374,711	372,050	356,469	338,608	306,305	301,745	343,928	337,087	335,327	345,448
Cement	260,803	260,803	252,197	250,110	239,417	239,417	207,078	201,340	224,551	244,633	248,806	249,328	249,328
Lumber	1,358,705	1,358,705	1,277,183	1,320,661	1,332,890	1,244,574	922,561	699,733	866,854	1,100,551	1,120,932	1,154,899	1,319,303
Paint and paint materials	634,869	634,869	634,809	634,869	630,425	615,188	554,876	496,468	503,451	584,079	584,079	584,079	584,079
Plumbing and heating	281,213	281,213	253,935	253,654	236,500	242,687	238,750	207,535	199,661	204,442	178,008	197,693	208,098
Structural steel	148,868	148,868	140,978	141,722	146,040	129,962	123,709	120,434	123,560	135,023	136,810	141,425	168,870
Other building materials	1,390,395	1,390,395	1,345,902	1,338,950	1,338,950	1,277,773	1,234,671	1,155,418	1,188,788	1,268,040	1,245,794	1,262,479	1,309,752
Aggregate value		4,454,884	4,288,135	4,314,677	4,296,272	4,106,070	3,620,253	3,187,233	3,408,610	3,880,696	3,851,516	3,926,230	4,184,878
Index number (per cent of 1926)		100 0	96 3	96 9	96 4	92 2	81 3	71 5	76 5	87 1	86 5	88 1	93 9

Source. See Table 136. Figures for 1927–1937 are derived from p_0q_0 column and relatives of Table 142

606

results sometimes attending the use of haphazard weighting. It may seem surprising that the simple arithmetic average obtains, in this case, better results than the index with base year weights. Ordinarily this would not be the case if a reasonably large sample were selected. But in the present instance, using only one commodity for each subgroup, it so happened that giving equal weight to each commodity gave more accurate representation to each subgroup than did weighting each commodity in accordance with its own importance. Generally speaking, indexes involving the simple arithmetic average of relatives or the simple aggregates would be the least desirable.

Quantity Index Numbers

Aggregative type. An aggregative index number of quantity (physical volume) is the counterpart of the corresponding price index. The general quantity formula is

$$Q = \frac{\Sigma q_n p}{\Sigma q_o p}.$$

Taking the building materials data, the construction of simple aggregative quantity index numbers is illustrated by Table 145. With 1926 prices as weights, an index of the quantity of building materials marketed may be constructed as in Table 146. This index number is not an estimate of the physical volume of building construction, nor is the corresponding price index an index of the cost of building construction. A considerable amount of construction material is purchased for the purpose of maintenance and repairs rather than new construction; furthermore, no account has been taken of labor cost. Although our price index parallels fairly closely the official index for the price of building materials, we have no basis of comparison for our quantity indexes. Nor can great accuracy be claimed for them; in absence of complete data much of the quantity material from which the indexes were constructed was obtained by very rough estimates.

Just as the aggregative index number of price measures the changing value of a fixed aggregate of goods, so the aggregative index number of physical volume measures the changing value of a varying aggregate of goods at fixed prices. The price index answers the question: If we buy the same assortment of goods each year, but at *different prices*, how much will we spend each year? The physical volume index answers the question: If we buy *varying quantities* of specified goods each year, but at the same price, how much will we spend each year? While in the former case the difference in amount spent was due to price change, in the latter case the difference must, of course, be attributed to changes in quantities bought and sold, since prices were held constant.

A number of different methods of weighting are available for the con-

TABLE 145

CONSTRUCTION OF SIMPLE AGGREGATIVE INDEX NUMBERS OF QUANTITY OF BUILDING MATERIALS MARKETED, 1926–1937

(Quantities in thousands)

Commodity	Unit	1926 q_{26}	1927 q_{27}	1928 q_{28}	1929 q_{29}	1930 q_{30}	1931 q_{31}	1932 q_{32}	1933 q_{33}	1934 q_{34}	1935 q_{35}	1936 q_{36}	1937 q_{37}
Common building brick.	1,000	7,424	6,967	9,498	5,405	3,586	2,9.5	1,000	1,007	1,085	1,7.0	3,223	3,403
Portland cement.	barrel	149,513	150,766	160,540	155,121	146,702	113,2.0	60,792	57,754	70,779	69,0.2	102,497	106,116
Hard maple, No 1	1,000 board feet	882	823	840	876	610	19	222	235	370	11.	522	559
Outside white gloss house paint	gallon	39,710	11,012	12,419	46,297	40,370	33,0..	25,116	28,255	35,408	42,..0	48,562	51,583
Lavatories	each	1,772	1,892	1,810	1,902	1,197	3,2.8	629	440	511	..2	2,114	3,347
Structural steel	100 pounds	76,031	74,986	91,218	92,948	65,007	41,1.1	19,464	21,304	25,889	35,305	49,025	51,876
Building gravel	ton	57,796	61,090	61,158	71,829	57,189	42,.19	32,897	37,648	41,116	44,2.9	48,815	51,831
Aggregate amount		333,188	343,136	372,110	374,378	314,997	233,536	149,120	146,643	175,158	194,931	254,758	268,715
Index number (per cent of 1926)		100 0	103 0	111 7	112 4	94 5	70 1	44 8	44 0	52 6	58 5	76 5	80 6

Source. Data are estimates, based upon data in different volumes of *Census of Manufactures*, *Wholesale Prices*, and other publications

TABLE 146

CONSTRUCTION OF 1926–1937 AGGREGATIVE INDEX NUMBERS OF QUANTITY OF BUILDING MATERIALS MARKETED, WEIGHTED BY 1926 PRICES

(Values in thousands of dollars)

Commodity	1926 price per unit p_{26}	Value of quantities marketed in specific years at 1926 prices											
		1926 $q_{26}p_{26}$	1927 $q_{27}p_{26}$	1928 $q_{28}p_{26}$	1929 $q_{29}p_{26}$	1930 $q_{30}p_{26}$	1931 $q_{31}p_{26}$	1932 $q_{32}p_{26}$	1933 $q_{33}p_{26}$	1934 $q_{34}p_{26}$	1935 $q_{35}p_{26}$	1936 $q_{36}p_{26}$	1937 $q_{37}p_{26}$
Common building brick.	$11.91	104,290	96,912	90,479	75,200	49,892	31,805	13,913	14,010	15,096	24,890	44,812	47,346
Portland cement.	1.71	240,803	273,140	279,685	270,531	275,848	197,151	121,717	100,7.3	123,439	121,508	178,755	185,076
Hard maple, No 1	65.67	19,104	45,59	43,469	48,770	.,631	19,1.0	12,359	13,0.3	20,599	24,218	29,061	31,121
Outside white gloss house paint	2.208	87,716	90,31	94,765	102,231	90,150	72,917	55,456	62,.87	78,181	94,677	107,225	113,895
Lavatories	12.171	21,927	23,112	22,763	23,5.5	18,524	15,319	7,783	5,115	6,323	10,666	26,159	41,416
Structural steel	1.958	118,869	117,15	176,617	181,402	127,284	80,613	38,111	41,.3	50,691	68,540	95,991	101,573
Building gravel	.911	51,386	55,7.5	65,488	67,501	.,815	39,756	30,956	35,1.7	38,690	41,976	45,945	48,773
Aggregate value		726,125	733,642	773,001	769,843	630,144	457,294	280,295	272,788	333,019	386,175	527,968	569,190
Index number (per cent of 1926)		100 0	101 0	106 5	106 0	86 8	63 0	38 6	37 6	45 9	53 2	72 7	78 4

Source. Tables 138 and 145

struction of quantity index numbers, and in general the same considera-
tions apply that were discussed in connection with price index numbers.
In obtaining price weights which are averages of two or more years, the
average prices should be weighted average prices, obtained by dividing
the total value sold in these years by the total number of units in those
same years.   Thus, if average quantities of base and given years are used,
we have the rather formidable looking formula

$$Q = \frac{\Sigma q_n\left(\dfrac{p_o q_o + p_n q_n}{q_o + q_n}\right)}{\Sigma q_o\left(\dfrac{p_o q_o + p_n q_n}{q_o + q_n}\right)} = \frac{\Sigma q_n\left(\dfrac{v_o + v_n}{q_o + q_n}\right)}{\Sigma q_o\left(\dfrac{v_o + v_n}{q_o + q_n}\right)}.$$

Likewise, if the common factor method is used, the price weight should
be derived from the largest value that is common to all the years in
question.

Averages of relatives.   This method of constructing quantity index
numbers is strictly analogous to the method applied to the measuring of
price changes.   The procedure is illustrated by Tables 147 and 148.   As
was found to be true with price index numbers, the use of base year value
weights produces the same result as the aggregative method employing
base year quantity weights.

Although, whenever it is applicable, the aggregative method is to be
preferred to the average of relatives method on account of ease of compu-
tation and simplicity of meaning, there are circumstances when the aggre-
gative method cannot be used.   When the relatives which are to be aver-
aged are percentages, not of a fixed base but of a changing normal, the
average of relatives method is necessary.   In other words the aggregative
method cannot be used if an index of business cycles is to be constructed,
since the data to be averaged are percentages of trend and seasonal.   The
Federal Reserve Bank of New York Monthly Index of Production and
Trade, described in the next chapter, is an illustration of this point.

Usually the weights selected for an average of quantity relatives are in
proportion to the values in exchange of the different series.   Occasionally,
some consideration is given also to the relative amplitude of the different
series, if they are cyclical relatives.   Several illustrations of this technique
will likewise be given in Chapter XXI.   If an index is constructed, not
for the purpose of *measuring* changes but for the purpose of *forecasting*
changes, the basis of selecting will be not the economic importance of the
different series represented, but their importance for purposes of forecast-
ing.   See description of the Bradford B. Smith Forecasting Index on pp.
817–820.

Chapter XXI will describe methods of constructing a number of im-

TABLE 147

CONSTRUCTION OF INDEX NUMBERS OF QUANTITY OF BUILDING MATERIALS MARKETED BY SIMPLE ARITHMETIC MEANS OF QUANTITY RELATIVES, 1926–1937

Commodity	1926	1927	1928	1929	1930	1931	1932	1933	1934	1935	1936	1937
Common building brick	100.0	93.8	87.5	72.8	48.3	30.8	13.5	13.6	14.6	24.1	43.4	45.8
Portland cement	100.0	104.8	107.1	103.7	98.1	75.7	46.7	38.6	47.3	46.6	68.5	71.0
Hard maple No. 1	100.0	93.3	88.3	99.3	72.6	39.6	25.2	26.6	42.0	49.3	59.2	63.4
Outside white gloss house paint	100.0	103.2	108.0	116.5	101.6	83.1	63.2	71.1	89.1	107.9	122.2	129.8
Lavatories	100.0	106.8	103.8	107.3	84.5	69.9	35.5	24.8	28.8	48.6	119.3	188.9
Structural steel	100.0	98.1	118.7	122.3	85.5	54.2	25.6	28.0	34.1	46.0	64.5	68.2
Building gravel	100.0	105.7	120.2	124.3	98.9	73.1	56.9	65.1	71.1	76.6	84.5	89.7
Average	100.0	100.8	104.8	106.6	84.2	60.9	38.1	38.3	46.7	57.0	80.2	93.8

Source: Table 145.

610

TABLE 148

CONSTRUCTION OF INDEX NUMBERS OF BUILDING MATERIALS MARKETED BY ARITHMETIC MEANS OF QUANTITY RELATIVES, WEIGHTED BY BASE YEAR VALUES, 1926–1937

(Values in thousands of dollars)

Commodity	1926 value p_0q_0	1926 value multiplied by quantity relative of specified year											
		1926	1927	1928	1929	1930	1931	1932	1933	1934	1935	1936	1937
Common building brick	103,286	103,286	96,882	90,375	75,192	49,887	31,812	13,944	14,047	15,080	24,892	44,826	47,305
Portland cement	260,803	260,803	273,322	279,581	270,453	255,848	197,428	121,795	100,670	123,360	121,534	178,650	185,170
Hard maple No. 1	49,104	49,104	45,814	43,359	48,760	35,650	19,445	12,374	13,062	20,624	24,208	29,070	31,132
Outside white gloss house paint	87,746	87,746	90,554	94,766	102,224	89,150	72,917	55,455	62,387	78,182	94,678	107,226	113,894
Lavatories	21,927	21,927	23,418	22,760	23,528	18,528	15,327	7,784	5,438	6,315	10,657	26,159	41,420
Structural steel	148,868	148,868	146,040	176,706	182,066	127,282	80,686	38,110	41,683	50,764	68,479	96,020	101,528
Building gravel	54,386	54,386	57,486	65,372	67,602	53,788	39,756	30,946	35,405	38,668	41,660	45,956	48,784
Aggregate value		726,120	733,516	772,919	769,825	630,133	457,371	280,408	272,692	332,993	386,108	527,907	569,233
Per cent of 1926.		100.0	101.0	106.4	106.0	86.8	63.0	38.6	37.6	45.9	53.2	72.7	78.4

Source: p_0q_0 values are from Table 143.   Figures for 1927–1937 are derived from p_0q_0 column and relatives of Table 147

portant indexes, and will discuss various points of technique and theory not covered in this chapter.

Selected References

R. W. Burgess *Introduction to the Mathematics of Statistics,* Chapter VI; Houghton Mifflin Co., Boston, 1927.

W. L. Crum, A. C. Patton, and A. R. Tebbutt: *Introduction to Economic Statistics,* Chapter XVIII; McGraw-Hill Book Co., New York, 1938.

F. C. Mills: *Statistical Methods Applied to Economics and Business* (Revised Edition), pages 161–199; Henry Holt and Co., New York, 1938.   A discussion of price indexes.

F. C Mills *The Behavior of Prices;* National Bureau of Economic Research, New York, 1927.   Not a treatise on index numbers but a study of the raw materials of price index numbers.

W. C. Mitchell: "The Making and Using of Index Numbers," *Bulletin 656 of the U. S. Bureau of Labor Statistics.*   A reprint of Part I of Bulletin No. 284. This pamphlet should be read by every student of index numbers.

J. R. Rigglemann and I. N. Frisbee: *Business Statistics* (Second Edition), Chapter IX; McGraw-Hill Book Co., 1938.

CHAPTER XXI

INDEX NUMBER THEORY AND PRACTICE

The object of this chapter is two-fold. First, the theory of index numbers and certain refinements of technique will be further discussed. Second, a description of a number of indexes will be given. The indexes were selected partly on account of their wide usefulness, and partly on account of the interesting technique which they employ. In general it will be found that in actual practice none of the procedures outlined in Chapter XX will be followed exactly, but that in each case there will be circumstances which justify special modifications of method.

Index Number Concepts

Mathematical tests. One school of thought on index numbers believes that there may be such a thing as a perfect index number formula, and that such a formula can be recognized by its ability to meet certain mathematical tests of consistency. Whether or not those tests are logically valid is an open question. Not only can an index be considered "ideal" if it meets those tests, according to this theory, but other indexes that do not meet them can be graded according to how closely they approximate them in actual practice.

The tests are derived by the logic of analogy. Anything that is true of an individual commodity should also be true of a group of commodities considered as a whole. If a pound of cotton was worth 125 per cent as much in 1936 as it was in 1926, then the 1926 price was 80 per cent of the 1936 price. Reasoning by analogy, if an index number for 1936 was 125 with respect to a 1926 base, then a similar index number for 1926 should be 80 with respect to a 1936 base. In other words, an index number should work backward as well as forward. Again, suppose that a commodity increases from 40 cents to 60 cents and that the sales increase from 2 units to 4 units. The price is 150 per cent of the base year, the quantity sales are 200 per cent, while the value is $1.50 \times 2.00 = 3.00$ times the base year, or 300 per cent of the base year. This is verified by

612

noting that $\dfrac{.60 \times 4}{.40 \times 2} = 3$. Once more reasoning from analogy, it may be argued that a price index times a quantity index computed from the same data should equal the relative value of the transactions in the given year with respect to the base year.   In other words, if

$$\frac{p_n}{p_o} \times \frac{q_n}{q_o} = \frac{p_n\,q_n}{p_o\,q_o} = \frac{v_n}{v_o},$$

then it should be true that

$$P \times Q = \frac{\Sigma p_n q_n}{\Sigma p_o q_o} = V.$$

As indicated in the preceding paragraph, there are two tests which are considered especially important by the "mathematical test" school.   These may be called: (1) the *time* reversal test; (2) the *factor* reversal test.

The time reversal test may be stated more precisely as follows: If the time subscripts of a price (or quantity) index number formula be interchanged, the resulting price (or quantity) formula should be the reciprocal of the original formula. If we take the formula

$$\frac{\Sigma p_n q_o}{\Sigma p_o q_o}$$

and interchange the time subscripts, the resulting formula is

$$\frac{\Sigma p_o q_n}{\Sigma p_n q_n}.$$

But

$$\frac{\Sigma p_n q_o}{\Sigma p_o q_o} \times \frac{\Sigma p_o q_n}{\Sigma p_n q_n} \neq 1;$$

hence the test is not met.   On the other hand, the formula

$$\sqrt{\frac{\Sigma p_n q_o}{\Sigma p_o q_o} \times \frac{\Sigma p_n q_n}{\Sigma p_o q_n}}$$

becomes

$$\sqrt{\frac{\Sigma p_o q_n}{\Sigma p_n q_n} \times \frac{\Sigma p_o q_o}{\Sigma p_n q_o}},$$

the product of the two expressions is unity, and the "ideal" index meets the time reversal test.

The factor reversal test may be stated in this way: If the p and q factors in a price (or quantity) index formula be interchanged, so that a quantity

(or price) index formula is obtained, the product of the two indexes should give the true value ratio

$$\frac{\Sigma p_n q_n}{\Sigma p_o q_o}.$$

Again taking the formula

$$\frac{\Sigma p_n q_o}{\Sigma p_o q_o},$$

we transform it into

$$\frac{\Sigma q_n p_o}{\Sigma q_o p_o}.$$

This is a quantity index, but since

$$\frac{\Sigma p_n q_o}{\Sigma p_o q_o} \times \frac{\Sigma q_n p_o}{\Sigma q_o p_o} \neq \frac{\Sigma p_n q_n}{\Sigma p_o q_o},$$

the test is not met.   However, we find that

$$\sqrt{\frac{\Sigma p_n q_o}{\Sigma p_o q_o} \times \frac{\Sigma p_n q_n}{\Sigma p_o q_n}}$$

transforms into

$$\sqrt{\frac{\Sigma q_n p_o}{\Sigma q_o p_o} \times \frac{\Sigma q_n p_n}{\Sigma q_o p_n}}.$$

The product of these two "ideal" indexes is

$$\frac{\Sigma p_n q_n}{\Sigma p_o q_o},$$

and the test is met.

The "ideal" index number is so called because it is one of an extremely limited number of indexes that meet both of these tests.

Relationship of formula to use.   The concept of an "ideal" index is attacked by index number students belonging to a different school of thought on the ground that the analyst cannot say exactly what the "ideal" index measures; he can only assert vaguely that it measures a change in the price level, or use some similar expression.   To Willford I. King[1] the logical procedure is to ask a specific question, and then to devise a formula which will answer that specific question.   For instance, the formula $\frac{\Sigma p_n q_o}{\Sigma p_o q_o}$ compares the cost in the present year with the cost in the base year of supporting the physical scale of living which obtained in the base

---

[1] See Willford I. King, *Index Numbers Elucidated*, Longmans, Green and Company, New York, 1930, especially Chapter III.

year.   While this is a specific question, it may not be the most useful question to ask.   Just what is an appropriate question to ask is an important problem facing the person conducting the investigation.   In Chapter XX Keynes was interpreted as believing it appropriate that, for measuring changes in the value of money, one should first seek an index number that would measure the changing cost of aggregates of goods yielding the same utility to similar groups of persons at two periods.

Now the formula $\dfrac{\Sigma p_n q_o}{\Sigma p_o q_o}$ assumes that, if their tastes do not change, people will continue to buy the same amounts of goods no matter how great the price rise or fall, while actually there is a shift from those items which are becoming more expensive to those which are becoming cheaper.   From a Keynesian point of view, then, this formula would have an upward "bias," since the cost of obtaining the same quantity of goods would be higher than the cost of obtaining the same quantity of utility.

The formula $\dfrac{\Sigma p_n q_n}{\Sigma p_o q_n}$, on the other hand, compares the cost of supporting one's present physical scale of living with its cost in the base year.   This formula, from the same point of view, has a downward "bias," since no sensible person would have bought the same goods in the base year as he does now (even granting the same tastes and environment) because the relative prices of goods would have been different.   The cost of obtaining the present year's bill of goods in the base year would have been greater than the cost of obtaining the current year's economic satisfactions.

Fisher's "ideal" index formula is the geometric mean of two index numbers biased (or inappropriate) in opposite directions; and many persons hold that the average of two wrong answers does not necessarily give one right answer, even though the two errors are in opposite directions and even though the formula is internally consistent.   On the other hand, it is doubtful that Keynes' common factor method will in actual practice answer Keynes' question any better than (if as well as) the "ideal" index number.   Changes in relative prices with consequent changes in relative quantities purchased may reduce the value of the common factor to a small proportion of the total goods bought.   Nevertheless, it is a meritorious attempt to arrive at a logical decision as to exactly what one is trying to measure.

For purposes of measuring changes in the value of money (purchasing power of the dollar), it is customary to use the reciprocal of a price index. Ferger, however, argues that this is illogical.[2]   Just as a price index aver-

---

[2] See Wirth F. Ferger, "Distinctive Concepts of Price and Purchasing Power Index Numbers," *Journal of the American Statistical Association*, Vol XXXII. June 1936, pp. 258–272.

ages together price changes of specific commodities, so a purchasing power index should average together changes in the purchasing power of the dollar for specific commodities. If the price of corn is $.50 per bushel, the purchasing power of the dollar for corn is 2 bushels. Designating units of purchasing power per dollar by the symbol u, Ferger suggests this purchasing power index number formula:

$$\text{Purchasing power} = \frac{\Sigma\left(\dfrac{u_n}{u_o}\,p_o q_o\right)}{\Sigma p_o q_o}.$$

But since $u = \dfrac{1}{p}$, we may write

$$\text{Purchasing power} = \frac{\Sigma\left(\dfrac{p_o}{p_n}\,p_o q_o\right)}{\Sigma p_o q_o}.$$

This expression is the reciprocal of the harmonic mean of price relatives weighted by base year values, since the latter is

$$1 \div \frac{\Sigma\left(\dfrac{1}{p_n \div p_o}\,p_o q_o\right)}{\Sigma p_o q_o} = 1 \div \frac{\Sigma\left(\dfrac{p_o}{p_n}\,p_o q_o\right)}{\Sigma p_o q_o} = \frac{\Sigma p_o q_o}{\Sigma\left(\dfrac{p_o}{p_n}\,p_o q_o\right)}.$$

So Ferger's formula is still in effect (though not in concept) the reciprocal of a price index, though not the usual index based on the arithmetic mean. Presumably it would be possible to alter somewhat the weighting system without doing violence to his concept.

If we accept the idea that the purpose of an index number determines its formula, we need not, necessarily, abandon the "ideal" formula. It would be possible to maintain that, although the formula is not a perfect solution to every index number problem, nevertheless there are purposes for which it is especially suited, as for instance the analysis of value changes into constituent price changes and quantity changes. However, it seemingly would have to be abandoned as a theoretically sound index if we take the position that every index number must answer a specific question couched in layman's English.

The Chain Index

Federal Reserve Bank of New York Index of Trend of Production and Trade in the United States. This index runs from 1830 to date, and is based upon 1870 as 100. The basic index (that is, the index before trend is computed) is on a 5-year rather than an annual basis, since it was impossible to obtain satisfactory data for shorter periods. Some series could

be obtained only at 10-year intervals, in which case it was necessary to make an estimate for intervening 5-year intervals.   The list of series available for inclusion in the index was small in 1830, but was much larger a century later:

1830	*1930*
Sugar imports	Crop production
Cotton consumption	Boots and shoes
Newspapers, etc., published	Sugar imports
Coal	Rubber imports
Lead	Cotton consumption
Pig iron	Wool consumption
Volume of U. S imports	Silk imports
U S tonnage cleared	Lumber cut
N Y State canal traffic	Newspapers, etc., published
Gross postal receipts	Coal
	Lead
	Copper
	Zinc
	Petroleum
	Cement
	Common brick
	Face brick
	Window glass
	Pig iron
	Steel
	Motor vehicles, passenger and truck
	Gas (manufactured) sold
	Electricity
	Employment in manufacturing
	Postage stamps issued
	Railway freight carried
	Telephones
	Trade and transportation employment

With a few exceptions these series are in physical, rather than value, units.   Not only is the 1930 list longer than the 1830 list, but four of the ten 1830 series have disappeared from the index.   In view of the growing and changing list of items it was impossible to compare each year directly with 1870.   Consequently, the percentage that each year was of the year five years preceding was computed.   Furthermore, since the sample was of necessity unsatisfactory from the standpoint of size, representativeness, and accuracy—particularly during the early years—and the dispersion of the relatives so great, it was thought advisable not to average together all the items each year, but to compute a modified mean from the central three or four items.   The procedure for five successive pairs of years is shown in Table 149.   The modified means appearing in the bottom row of the table may be called link relative index numbers.

The basic index is now obtained by a process of successive multiplica-

TABLE 149

COMPUTATION OF MODIFIED MEANS OF PERCENTAGES OF FIFTH PRECEDING YEAR FOR SELECTED PRODUCTION AND TRADE SERIES, 1860-1885

(In some instances the same rate of growth is recorded for two successive periods This is usually because those series were available by 10-year intervals only, and the 5-year interval was interpolated at half the 10-year rate   Railway freight through 1870 was in units of ton miles, beginning 1875 it was in units of tons carried Because of lack of comparability this item had to be omitted for the period 1870-1875 If we should estimate this series at about 200 for the period, the effect would be to exclude rubber imports from the positional mean for 1870-1875 and raise the average to 134 3)

1860-1865

Series	Per cent
Petroleum	499 5
Rubber imports	258 5
Wool consumption	217 2
Postage stamps issued	179 0
Railway freight carried	174 6
Gross postal receipts	171 8
Coal	163 1
Lumber cut	128 1
Employment in manufacturing	125 2
Crop production index	125 1
Trade and transportation employment	120 6
Newspapers, etc, published	120 4
Copper	118 1
N Y State canal traffic	101 7
Pig iron	101 3
Lead	94 2
Sugar imports	93 8
Volume of U S imports	87 0
U S tonnage cleared	75 3
Cotton consumption	73 0
Modified mean	122 8

1865-1870

Series	Per cent
Rubber imports	258 5
Railway freight carried	227 6
Petroleum	217 0
Silk imports	202 8
Pig iron	200 1
Sugar imports	183 6
Cotton consumption	167 0
Copper	148 2
Crop production index	143 9
Coal	139 2
U S tonnage cleared	138 5
Gross postal receipts	135 6
N Y State canal traffic	130 5
Lumber cut	128 1
Employment in manufacturing	125 2
Lead	121 1
Postage stamps issued	120 8
Trade and transportation employment	120 6
Newspapers, etc, published	120 4
Wool consumption	113 0
Volume of U S imports	87 0
Modified mean	137 8

1870-1875

Series	Per cent
Steel	565 2
Lead	334 8
Petroleum	232 8
Silk imports	188 7
Telegraph messages	187 3
Coal	162 1
Sugar imports	150 0
Postage stamps issued	145 7
Copper	142 9
Newspapers, etc, published	134 0
Boots and shoes manufactured	126 0
Rubber imports	125 0
Crop production index	124 8
Trade and transportation employment	123 1
Pig iron	122 0
Cotton consumption	122 0
Horsepower in manufacturing	121 0
Lumber cut	119 0
Employment in manufacturing	115 3
Wool consumption	110 9
Modified mean	132 0

1875-1880

Series	Per cent
Steel	549 7
Silk imports	242 5
Petroleum	205 5
Telephones, number	190 0
Pig iron	190 0
Railway freight carried	170 8
Telephone messages	170 3
Coal	160 6
Sugar imports	153 7
Lead	154 0
Copper	150 0
Cotton consumption	149 0
Zinc	147 0
Rubber imports	140 0
Coal	136 2
Crop production index	128 8
Postage stamps issued	128 4
Boots and shoes manufactured	126 0
Newspapers, etc, published	123 5
Trade and transportation employment	123 1
Horsepower in manufacturing	121 0
Lumber cut	119 0
Employment in manufacturing	121 0
Sugar imports	115 3
Modified mean	145 3

1880-1885

Series	Per cent
Copper	274 4
Cement	200 2
Telephones, number	200 0
Zinc	179 0
Silk imports	168 2
Postage stamps issued	167 7
Coal	154 7
Sugar imports	149 0
Railway freight carried	147 4
Telegraph messages	144 1
Rubber imports	144 0
Newspapers, etc, published	138 8
Steel	137 3
Trade and transportation employment	133 3
Horsepower in manufacturing	132 0
Lead	131 9
Window glass	124 9
Employment in manufacturing	
Lumber cut	124 6
Boots and shoes manufactured	115 0
Cotton consumption	112 0
Crop production index	109 5
Wool consumption	105 2
Pig iron	105 0
Petroleum	83 2
Modified mean	136 3*

Source: Research Department, Federal Reserve Bank of New York.   * Modified means computed from central items rounded to whole per cents.

tion, beginning with 1870 as 100, as illustrated by Table 150.   The logic
is as follows:

> Let 1870 be the base year, or 1870 = 100.0;
> 1875 is 132 0 per cent of 1870, or 1.320 × 100.0 = 132.0;
> 1880 is 145 3 per cent of 1875, or 1 453 × 132 0 = 191.8;
> 1885 is 136.3 per cent of 1880, or 1 363 × 191 8 = 261.4; and so on.

To obtain years preceding 1870, we must express each year as a percentage
of the year following.   Thus, if 1870 is 137.8 per cent of 1865, then 1865 is
$\frac{1.000}{1.378} = 72.57$ per cent of 1870.   This is the last number given in column
4 of Table 150.   The other numbers in this column were obtained in a
similar fashion.   We may then proceed to reason as follows:

> Let 1870 be the base year, or 1870 = 100 0,
> 1865 is 72 57 per cent of 1870, or 7257 × 100 = 72 57;
> 1860 is 81.43 per cent of 1865, or 8143 × 72.57 = 59 09;
> 1855 is 81 37 per cent of 1860, or .8137 × 59 09 = 48.08, and so on.

TABLE 150

CONSTRUCTION OF CHAIN INDEX OF PRODUCTION AND TRADE OF THE UNITED STATES
(FROM LINK RELATIVE INDEX NUMBERS) AND INDEX OF TREND, 1830–1935

Period (1)	Link relative, later year as percentage of earlier (2)	Period (3)	Link relative, earlier year as percentage of later (4)	Year (5)	Basic index (6)	Index of trend (7)
1830–1835	147.1	1835–1830	67 98	1830	10 80	11 4
1835–1840	124.8	1840–1835	80 13	1835	15 88	15 1
1840–1845	135.1	1845–1840	74 02	1840	19 82	20 1
1845–1850	140 2	1850–1845	71 33	1845	26 77	26.7
1850–1855	128 1	1855–1850	78.06	1850	37 53	35 5
1855–1860	122 9	1860–1855	81 37	1855	48 08	47 1
1860–1865	122 8	1865–1860	81 43	1860	59 09	63.4
1865–1870	137 8	1870–1865	72.57	1865	72 57	84 5
. .	.	.	. .	1870	100 0	112 0
1870–1875	132 0	. . .	. . .	1875	132.0	147.3
1875–1880	145.3	. .	. .	1880	191 8	192 5
1880–1885	136.3	. .	. . .	1885	261.4	249.8
1885–1890	132.7	. . .	. . .	1890	346.9	322 0
1890–1995	121.7	. . .	. . .	1895	422 2	412.2
1895–1900	129.3	. . .	. . .	1900	545 9	524.0
1900–1905	131 8	. . .	. . .	1905	719.5	661 5
1905–1910	123.7	. . .	. .	1910	890 0	829.5
1910–1915	119.75	. .	. . .	1915	1,065.8	1,032.9
1915–1920	121.5	. . .	. . .	1920	1,294 9	1,277.5
1920–1925	129.0	. . .	. . .	1925	1,670.4	1,569.1
1925–1930	94.5	. .	. . .	1930	1,578.5	1,914.0
. . .	. . .	. . .	. . .	1935	. .	2,318.9

Source: Table 149 and Research Department, Federal Reserve Bank of New York

In Table 150 there is an additional column for the trend. It was recognized that complete accuracy could not be expected for each individual index number; but it was believed that a fitted trend would smooth out the errors and give a reasonably accurate picture of the nature of the growth of United States physical production and trade. The trend equation fitted to the 1830–1860 data is

$$\log Y_C = .42677 + .0246669X,$$

with origin at 1845 and X units of one year. The trend values from this
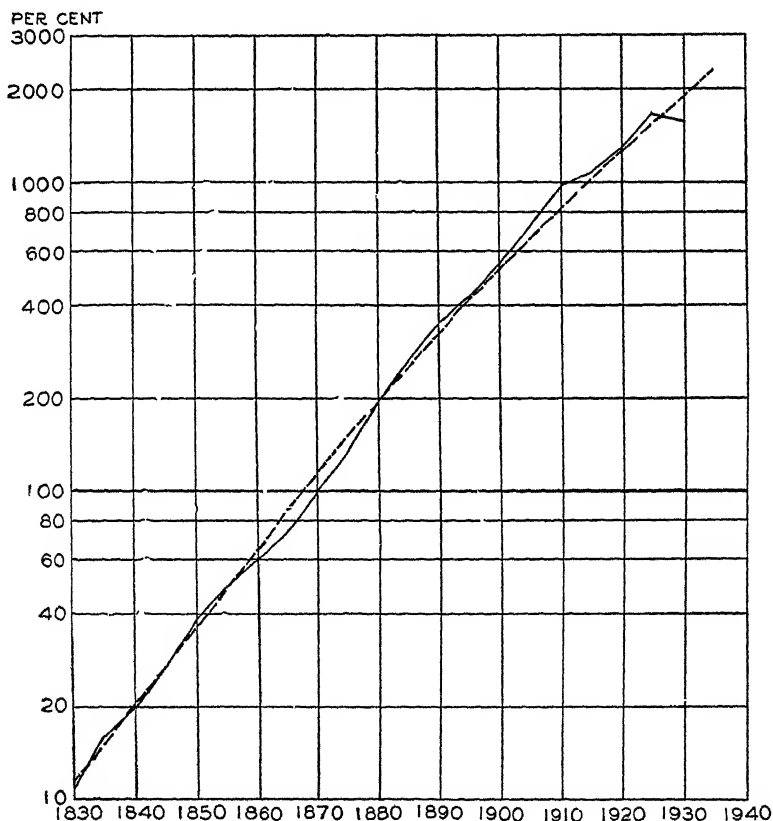
Chart 214.   Federal Reserve Bank of New York Annual Index of Production and Trade and Fitted Trend, 1870-1930.   (1870 = 100.   Data of Table 150.)

equation are for 1830–1855. For trend values from 1856 to date, the following equation was obtained by fitting to the 1840–1930 data:

$$\log Y_C = -.011585 + .0289066X - .0000597117X^2.$$

The origin for this equation is 1830. Trend values are computed for each

year, though Table 150 shows trend values for each fifth year only. The index and trend are also shown in Chart 214. This index is not published in any periodical, but may be obtained upon request from the Federal Reserve Bank of New York.

The chief advantage of chain index numbers (described very briefly in Chapter XX also) is the ease with which new items may be added to the sample and old items dropped, or new items substituted for old ones. The link relative index numbers, of which the chain index is composed, compares data that are reasonably comparable, since the intervening period is short and the quality of the items included has not had much time to change. Furthermore, if the index is weighted, a weighting system may be had that compares the two years with substantial accuracy. But when the links are chained together, precise comparability is lost, and the meaning of the index cannot be stated in simple language.

Circular test. Among the tests sometimes advocated for index numbers is the circular test, which may be regarded as an extension of the time reversal test but which is applicable to chain indexes only. It is argued that, just as an index number formula should work backward as well as forward between two years, so it should work in circular fashion among a number of years. The working of the test is most easily explained by reference to an illustration such as that given by Table 151. In that

TABLE 151

CIRCULAR TEST APPLIED TO FEDERAL RESERVE BANK OF
NEW YORK INDEX OF PRODUCTION AND TRADE IN
THE UNITED STATES, 1860–1885

Period	Link relative	Year	Chain index
.	.	1870	100 00
1870–1875	132 0	1875	132 00
1875–1880	145 3	1880	191.80
1880–1885	136 3	1885	261 42
1885–1860	20 84*	1860	54.48
1860–1865	122 8	1865	66.90
1865–1870	137 8	1870	92 19

* Positional mean is average of

Sugar imports	25 47
Wool consumption	22 72
Pig iron (production)	20 38
Postage stamps issued	14 79

Source. Computed from data provided by Research Department, Federal Reserve Bank of New York

table, which uses the data of Table 150, a circular chain is forged, which begins with 1870, chains successively the links 1870–1875, 1875–1880,

1880–1885. 1885–1860, 1860–1865, 1865–1870, and arrives at the starting point 1870 with a new index number. The link 1885–1860 is the mean of the four central 1860 percentages with respect to 1885. As can be seen, there is a discrepancy between the two index numbers for 1870: 100.00 and 92.19.

The reasons why most chain indexes do not meet the circular test are as follows:

(1) The type of formula is not one which meets the test. There are very few formulae that do. Among those that do are the geometric mean with constant value weights, and the aggregative index with constant quantity (or price) weights. The simple arithmetic mean gives a terminal value which is always greater than 100. The reason for the discrepancy is the implicit change in the quantity element in the value weights for a price index (see p. 599), or in the price element in the value weights for a quantity index. The simple median has no bias one way or the other, but it is only by accident that this median may meet the test. A modified mean which includes a large proportion of the available items would presumably have an upward bias. However, when the proportion of items is small, as in our present illustration, the discrepancy might be in either direction. The variable nature of the commodities averaged by our different modified means is the most important reason for our present discrepancy. Below is the list of series used. Of the 21 series appearing during the six periods, 17 are different; not one appears throughout, and only one series (Newspapers, etc., published) appears as often as three times.

1860–1865	*1865–1870*	*1870–1875*
Employment in manufacturing	Coal (production)	Copper (production)
Crop production index	U. S. tonnage cleared	Newspapers, etc , published
Trade and transportation employment	Gross postal receipts	Boots and shoes manufactured
Newspapers, etc , published		Rubber imports
1875–1880	*1880–1885*	*1885–1860*
Cotton consumption	Steel (production)	Sugar imports
Zinc (production)	Trade and transportation employment	Wool consumption
		Pig iron (production)
Rubber imports	Newspapers, etc., published	Postage stamps issued

(2) Generally an important reason for using a chain index is that the relative importance of the different items changes with the passage of time, and the chain index is one device for putting such changes into effect. Such changes in weights are intentional. But, as explained in the preceding paragraph, they may be unintentional—the result of using an inappropriate type of weighting for the index number formula used

Whether the changes are intentional or unintentional, they cause the chain index to fail to meet the circular test.

(3) Usually also the list of commodities changes over time, and this is an important reason for employing the chain index. In the present instance there were a few changes in the list of commodities during the period 1860–1885.

Whenever there is a change in the commodities used or the weights given to them, there is a break in comparability. Consequently most authorities argue that there is no reason why a chain index *should* meet the circular test. If the commodities and weights remain the same, most of the index number formulae will meet the test, if they are used properly. But if the commodities and weights remain the same, there is no reason for using the chain index.

Substituting New Commodities and Changing Weights

The preceding chain index illustration concerned a simple average of relatives. The use of weights would have introduced no new problem. However, when an aggregative type of index is used, the problem of adjusting weights puts certain pitfalls in the path of the statistician.

We shall again utilize the building material price data with which we are familiar. Let us assume that it is desirable to substitute California redwood for hard maple beginning 1936. Consequently there are two sets of $p_{36}q_{26}$ values shown in Table 152—one for the old series, using hard maple; and a second for the new series, which uses California redwood instead. The product for 1937 ($p_{37}q_{26}$) uses California redwood. The old series shows that 1936 prices are 88.12 per cent of the 1926 prices, while the new series indicates that 1937 is 108.24 per cent of 1936. But since the 1936 index is 88.12, the 1937 index, *relative to 1926*, is 88.12 × 1.0824 = 95.38.

The pitfall into which we have stepped is that lumber is given additional weight by the substitution of redwood for maple. This is due to the fact that the price of redwood is considerably higher than that of maple, and since the same quantity weights were used in the new series as the old, that part of the value aggregate due to the lumber was greatly enhanced. Specifically the value of the lumber in thousands of dollars was increased from 1,154,839 to 1,482,604; or, we may say, its relative importance in the index was increased from $\frac{1,154,839}{3,925,443} = 29.4$ per cent to $\frac{1,482,604}{4,253,154} = 34.9$ per cent.

As might be suspected, the remedy for this difficulty is to decrease the quantity weight for the lumber representative in the same ratio that the price of maple is to the price of redwood. This ratio is $\frac{60.75C}{47.322} = 1.283578$,

TABLE 152

CONSTRUCTION OF WEIGHTED AGGREGATIVE CHAIN INDEX NUMBERS OF BUILDING MATERIAL PRICES, ILLUSTRATING SUBSTITUTION OF NEW COMMODITY WITHOUT COUNTERBALANCING CHANGE IN WEIGHTS, 1926, 1936, AND 1937

Commodity	Quantity weights (thousands) q_{26}	1926		1936			1937	
		Price (dollars) p_{26}	Product (thousands of dollars) $p_{26}q_{26}$	Price (dollars) p_{36}	Product, old series (thousands of dollars) $p_{36}q_{26}$	Product, new series (thousands of dollars) $p_{36}q_{26}$	Price (dollars) p_{37}	Product (thousands of dollars) $p_{37}q_{26}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Common building brick	27,315	13 913	380,034	12 313	336,330	336,330	12 048	329,091
Portland cement . .	149,543	1 744	260,803	1 667	249,288	249,288	1 667	249,288
{Hard maple, No. 1 . .	24,405	55.673	1,358,700	47.322	1,154,839		.	
{California redwood . .		. .	.	60.750		1,482,604	72.000	1,757,160
Outside white gloss house paint . . .	287,531	2 208	634,868	2 031	583,975	583,975	2 031	583,975
Lavatories . .	22,726	12 374	281,212	8.699	197,693	197,693	9 086	206,488
Structural steel . .	76,031	1 958	148,869	1 860	141,418	141,418	2 215	168,409
Building gravel . .	1,477,572	.941	1,390,395	854	1,261,846	1,261,846	.886	1,309,129
Total			4,454,881	. .	3,925,443	4,253,154		4,603,540
Link relative index number			100 00	. .	88 12	100 00		108 24
Chain index number			100 00		88.12			$88\,12 \times 1\,0824 = 95\,38$

Source: Table 138 and p 604   Units are shown in Table 138

and the new quantity weight is 24,405 thousands of board feet ÷ 1.283578 = 19,010.6 thousands of board feet.   The use of this quantity weight gives the same $p_{36}q_{26}$ value for the old series as for the new, as can be seen by inspection of columns 6 and 7 of Table 153.   This means that the importance of the lumber item is unchanged after making the substitution of redwood for maple.   Now, since redwood increased in price by a relatively large percentage between 1936 and 1937, the 1937 link relative with 1936 as base is smaller than in Table 152, which gave too much weight to the item.   The new link relative is shown by Table 153 to be 107.38, which gives a 1937 chain index number of 88.12 × 1.0738 = 94.62.   This index number can also be derived directly from the data of Table 153 by the expression

$$P_{37} = \frac{\Sigma p_{37}q'_{26}}{\Sigma p_{26}q_{26}} = \frac{4{,}215{,}143}{4{,}454{,}881} = 94.62,$$

where q' is a quantity multiplied by the appropriate correction factor. We see therefore that column 7 of Table 153 is not needed, and that it is not necessary to obtain link relatives and chain them to the base, if we merely make the appropriate adjustment in quantity weights whenever a new commodity is introduced.

The chain index is useful, not only when making substitutions in the list of commodities, but when making a change in the weights.   Suppose that new quantity weights (say 1935 quantities) are to be introduced in 1936.   The 1936 index number, which uses the 1926 weights will be unchanged at 88.12.   We shall, however, have different products for comparing 1937 with 1936.   For instance (using hypothetical data not shown in our tables), we may have

$$\frac{\Sigma p_{37}q_{35}}{\Sigma p_{36}q_{35}} = \frac{4{,}376{,}918}{4{,}091{,}765} = 106.97.$$

This would give a chain index value for 1937 of 88.12 × 1.0697 = 94.26.

Now if the index is to be carried forward with the new set of weights, it would be somewhat laborious to chain each link relative to the index number for the preceding year in order to obtain a chain index.   A much easier procedure is to adjust the base year aggregate in such a way that the same result can be obtained directly.   In the present instance we may obtain the adjusted base year aggregate as follows:

$$\Sigma p_{26}q'_{35} = \Sigma p_{36}\,q_{35} \div P_{36}$$
$$= 4{,}091{,}765 \div .8812$$
$$= 4{,}643{,}401.$$

TABLE 153

CONSTRUCTION OF WEIGHTED AGGREGATIVE CHAIN INDEX NUMBERS OF BUILDING MATERIAL PRICES, ILLUSTRATING SUBSTITUTION OF NEW COMMODITY WITH COUNTERBALANCING CHANGE IN WEIGHTS, 1926, 1936, AND 1937

Commodity	Quantity weights (thousands) q_{26}	1926		1936			1937	
		Price (dollars) p_{26}	Product (thousands of dollars) $p_{26}q_{26}$	Price (dollars) p_{36}	Product, old series (thousands of dollars) $p_{36}q_{26}$	Product, new series (thousands of dollars) $p_{36}q'_{26}$	Price (dollars) p_{37}	Product (thousands of dollars) $p_{37}q'_{26}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Common building brick	27,315	13.913	380,034	12 313	336,330	336,330	12 048	329,091
Portland cement	149,543	1.744	260,803	1 667	249,288	249,288	1 667	249,288
Hard maple, No 1.	24,405	55 673	1,358,700	47 322				
California redwood	19,010.6*			60.750	1,154,893	1,154,894	72.000	1,368,763
Outside white gloss house paint	287,531	2.208	634,868	2 031	583,975	583,975	2 031	583,975
Lavatories	22,726	12.374	281,212	8 699	197,693	197,693	9.086	206,488
Structural steel	76,031	1 958	148,869	1 860	141,418	141,418	2 215	163,409
Building gravel	1,477,572	941	1,390,395	.854	1,261,846	1,261,846	.886	1,309,129
Total			4,454,881		3,925,443	3,925,444		4,215,143
Link relative index number			100 00		88.12	100 00		107 38
Chain index number			100 00		88 12			94 62†

* $19{,}010\ 6 = \frac{47\ 322}{60\ 750}\ 24{,}405$

† $94\ 62 = 88\ 12 \times 1\ 0738$, or $\frac{4{,}215{,}143}{4{,}454{,}881}$

Source: Table 138 and page 604

It is obvious that our final index number for 1937 may be obtained as follows:

$$P_{37} = \frac{\Sigma p_{37}\, q_{35}}{\Sigma p_{26}\, q'_{35}}$$
$$= \frac{4,376,918}{4,643,401} = 94.26 \text{ per cent.}$$

Subsequent index numbers with 1926 as a base are now computed directly by the expression

$$P_n = \frac{\Sigma p_n q_{35}}{\Sigma p_{26}\, q'_{35}}.$$

The United States Bureau of Labor Statistics claims these advantages for the two techniques explained above[3] (which may be employed in combination as well as separately): (1) The relative importance of a commodity is unaffected by the substitution of one price series for another. (2) The making of special group indexes is facilitated, since the chaining back process is side-stepped.

Some Price Indexes

United States Bureau of Labor Statistics Index of Wholesale Commodity Prices. This index, kept up to date on an annual, monthly, and weekly basis, is probably the most widely used price index in existence.[4] It extends back on a monthly and annual basis through 1890, and on a weekly basis through 1931. In January 1931 the number of price series included in the index was increased from 550 to 784, and the calculations were revised on that basis back to and including 1926. At present 813 series are included. A feature of the index is that index numbers of several groups of commodities are published, as well as those of the 813 series as a whole. These groups are:

(1) Farm products.
(2) Foods.
(3) Hides and leather products.
(4) Textile products.
(5) Fuel and lighting.

---

[3] See "Revised Method of Calculation of the Wholesale Price Index of the United States Bureau of Labor Statistics," by Jesse M. Cutts and Samuel T Dennis, *Journal of the American Statistical Association*, Vol XXXII, December 1937, pp 663-674

[4] For index numbers of wholesale prices 1890-1926, see United States Bureau of Labor Statistics Bulletin No. 543, *Wholesale Prices, 1930,* for 1926-1931, see Bulletin No. 572, *Wholesale Prices, 1931.* Figures for subsequent periods are shown in a monthly pamphlet, *Wholesale Prices,* in the *Monthly Labor Review,* and in monthly and weekly official releases. Mimeographed tables covering the entire period are available on request from the United States Bureau of Labor Statistics. For further description of the index, see the article by Cutts and Dennis referred to in footnote 3.

(6)   Metals and metal products.
(7)   Building materials.
(8)   Chemicals.
(9)   House-furnishing goods.
(10)   Miscellaneous.

Each of these groups is further subdivided into a number of subgroups, and separate index numbers are computed for each.   Articles properly falling under more than one classification are so listed: thus, structural steel is included under building materials as well as under metals and metal products; and eggs are considered both as a farm product and as a food.   In the computation of the general index, however, there is no dupli‚ cation; each commodity is counted only once.   In addition to this classifi‚ cation according to the nature of the commodity, there are two other classi‚‚ fications.   The first is based on origin: (1) farm products; (2) non-agricul‚ tural commodities; (3) all commodities other than farm products and foods‚ The second is based on degree of manufacture: (1) raw materials; (2) semi‚ manufactured products; (3) finished products.

The index numbers are constructed by the aggregative method.   Be‚ ginning in 1934, current prices of the 784 commodities have been multiplied by the average quantities of these commodities marketed in 1929 and 1931 (1929, 1930, and 1931, in the case of farm products and agricultural commodities).   Prices of the same commodities in 1926 were also multi‚ plied by the average of 1929 and 1931 quantities.   The products are summed for each period, and are divided by the 1926 product-sum, in order to express the index number as a percentage of 1926.   For 1934 the index number formula therefore is

$$P_{34} = \frac{\Sigma p_{1934} q_{1929,1931}}{\Sigma p_{1926} q_{1929,1931}}.$$

The 1929, 1931 weights were not used in computing the index numbers for every year.   The system of weighting is as follows:

Period	Mean of quantities in:
1913–1914 inclusive .	1909 and 1914
1915–1919 inclusive .	1914 and 1919
1920–1921 inclusive .	1919 and 1921
1922–1923 inclusive. . .	1921 and 1923
1924–1929 inclusive	1923 and 1925
1930–1931 inclusive . .	1925 and 1927
1932–1933 inclusive .	1927 and 1929
Beginning 1934	1929 and 1931

The quantity weights are obtained from census reports and from other governmental and private sources.   Current weights cannot be used ex‚ cept for revisions of the index, but the weights are kept as nearly up to date as is practicable.

The introduction of new commodities or of different weights from time to time necessitated some method of splicing the index together so that comparability would be retained, and from 1908 until 1937 the index has been computed as a chain index. Until 1937, when a new series was introduced to replace an old one which was no longer satisfactory or available, a procedure was followed similar to that shown in Table 152. (Of course, no substitution was made which destroyed the continuity of any series to such an extent as that illustrated.) Because of the theoretical and practical difficulties, the technique was modified in 1937 and is now like that indicated at the end of the preceding section dealing with substituting new commodities and changing weights. The weekly, monthly, and annual indexes are calculated independently of each other, and as a result of early imperfections in technique the weekly indexes differ slightly from the monthly index. The revised method, however, will bring the two indexes into substantial conformity from 1937 on.

The United States Bureau of Labor Statistics also makes a retail food price index, and an index of the cost of living, of which the retail food price index is an important component. The cost of living index is described in the following paragraphs, while a brief description of the index of retail food prices will be found in Davenport and Scott, *An Index to Business Indexes*, pp. 69–70, Business Publications, Inc., Chicago, 1937.

Changes in cost of living: *United States Bureau of Labor Statistics Index.* Index numbers of cost of living are currently computed each month by the National Industrial Conference Board,[5] and several times each year (usually quarterly) by the United States Bureau of Labor Statistics. The United States Bureau of Labor Statistics computes separate indexes for each of 32 cities with population over 50,000, and for the United States as a whole. The total cost of living index numbers for each city and for the nation are themselves obtained by combining six group indexes:

(1) Food.
(2) Clothing.
(3) Rent.
(4) Fuel and light.
(5) House-furnishing goods.
(6) Miscellaneous goods and services.

These separate indexes for each city are computed by the aggregative method, the weights in each city being the average amount of goods purchased per family per year by wage earners and low-salaried workers in

---

[5] This index is published currently in the *Survey of Current Business*. A discussion of it will be found in Conference Board Research Staff, *Cost of Living in the United States, 1914–1936*, National Industrial Conference Board, New York, 1936.

that city. Frequently a weight applied to a particular commodity is not the amount of that particular commodity purchased, but is derived from the amount spent for a group of related commodities represented by the commodity in question. The different group indexes are chain indexes; this method is used because of the constant change in the form in which consumer goods are offered for sale. In combining the separate group indexes into a cost-of-living index for any city, the six index numbers are weighted by 1917–1919 expenditures for the city in question.

No attempt is made to compare the relative cost of living in the different cities, but for each city a comparison is made between its current cost of living and its cost in the base period, 1923–1925. In combining the different city indexes into a national index, the weights are in proportion to the population of the metropolitan areas where retail prices are collected plus that of adjacent large urban centers in which it is believed that prices move in a similar fashion. A national index is made for all items and for each of the six groups.[6]

Geographical variations in cost of living: *W. P. A. Index of Intercity Differences.* The computation of index numbers comparing the cost of living in different cities is much more difficult than making a comparison over time, because consuming habits vary so much geographically, and because the purchase of the same goods yields such a varying amount of satisfaction in different regions. For instance, an expenditure for fuel that is ample for heating a house one season in a city in Maine would suffice for a number of years in Birmingham, Alabama. Nevertheless, the Works Progress Administration attempted a comparison between the cost of living in 59 cities in March 1935.[7] Instead of using as weights average quantities purchased per family, synthetic budgets were constructed, one at the maintenance level and one at the emergency level, and two sets of index numbers were constructed. Although the same general budget was used for each city, variations were made for many of the factors. For example, the need for heating and refrigeration varies with climate, and the need for transportation varies with population and land area. Consequently, in order to obtain the same standard of living in different cities, it was necessary to vary the quantity of these items allowed in the different cities. Cost of refuse disposal and school attendance was a part of the cost of living in some cities but not in others, depending on the local

[6] For further description, see "Revision of Index of Cost of Goods Purchased by Wage Earners and Lower-Salaried Workers," by Faith M. Williams, Margaret H. Hogg, and Ewan Clague, in the *Monthly Labor Review*, Vol. 41, September 1935, pp. 819–837.

[7] See *Intercity Differences in Cost of Living in March 1935, 59 Cities*, a report by Margaret Loomis Stecher, Works Progress Administration, Division of Social Research, Washington, 1937.

laws.   Likewise the allowance for taxation was not uniform.   On the other hand, the cost of postage, telephone calls, and insurance was assumed to be the same everywhere.

This brief discussion is not a description of the method of constructing this index; it is intended merely to give some idea of the tremendous difficulties which such a study entails.

Snyder's Index of the General Price Level.   The Federal Reserve Bank of New York maintains currently an index of the general price level, which is a weighted arithmetic average of a number of component price indexes.[8] These series are:

Series	Weight
Retail food prices	10
Rents	5
Other cost-of-living items	10
Industrial commodities at wholesale	10
Farm prices at the farm	10
Transportation costs	5
Realty values	10
Security prices	10
Equipment and machinery prices	10
Hardware prices	3
Automobile prices	2
Composite wages	15
General price level	100

Originally this index was intended as a companion index to Snyder's Index of the Volume of Trade, and Snyder was able to find a number of interesting relationships between these two and other logically related series.   Since the volume of trade index is no longer so inclusive as the price index, comparisons involving the two indexes no longer have the same significance.

Indexes of Physical Volume of Production and Trade

Board of Governors of the Federal Reserve System Index of Industrial Production.   This index, which is published monthly in the *Federal Reserve Bulletin,* is a good illustration of a quantity index constructed by the aggregative method.   The general index is itself composed of two indexes: an index of manufactures, with 53 series combined into 13 separated indexes of the different industrial groups, in addition to a number of subgroups; and an index of minerals, with 8 series.   The base period of the

---

[8] The current index numbers of the general price level are published in the *Monthly Review of Credit and Business Conditions, Second Federal Reserve District.*   For a detailed description of this index, see "The Measure of the General Price Level," by Carl Snyder. in *The Review of Economic Statistics,* February 1928, pp. 40–52.

indexes, which are carried back to the beginning of 1919, is the 3-year average of 1923–1925.

Much difficulty was encountered in obtaining suitable series for the index of manufactures. The construction industry, for instance, is not directly represented. Furthermore, many series that are available are not strictly comparable over a long period of time. An index of physical volume which fails to take into consideration the constant refinements made on mechanical contrivances will underestimate the growth of such industries. Nevertheless, many branches of industry which cannot be represented directly have been given indirect representation through other series. Thus, "steel ingots fairly measure current movements in more advanced stages of steel manufacture and less closely represent the broader swings of manufacturing activity in industries making finished products from steel."[9] However, the series in the manufacturing index are said to represent, directly or indirectly, 80 per cent of all manufacturing industries.

The prices used as weight multipliers are derived from value figures as follows:

1. *Minerals.* The total value of a given mineral produced during the three years 1923–1925, as reported by the Geological Survey of the United States Bureau of Mines, is divided by the total quantity produced in those same years.

2. *Manufactures.* The total value added by manufacture in 1923, as reported by the Census of Manufactures, is divided by the appropriate quantity figure for that year. Value added by manufacture is taken instead of actual value, in order to avoid counting an item in both its raw and its manufactured state.

Figures for 1923 are used solely because 1925 values were not available at the time this index was undertaken. Strictly speaking, each series is weighted not according to its own value but according to the relative importance of all industries that it represents in the index. In a number of instances the weighting of the different series was somewhat arbitrary.

In Chapter XX it was stated that, in dealing with time series representing physical volume, it is often desirable first of all to eliminate from the series irregularities that are due to the varying number of calendar days or working days in each month. In the case of most series, therefore, the Federal Reserve Board first reduces its monthly figures to daily averages

---

[9] *Federal Reserve Bulletin,* March 1927, Vol. XIII, No. 3, p. 170. The writers are indebted to the Division of Research and Statistics, Board of Governors of the Federal Reserve System for part of the information concerning this index. Since publication of this text the index has been revised. The new base is the five-year average 1935–1939 The average of relatives method has replaced the aggregative method, and other changes have been made. See Federal Reserve Bulletin, Vol. XXVI, August 1940, pp. 753–771 and September 1940, pp. 912–923.

by dividing them by the appropriate number of working days in the month. It was found that the industries fell into three groups in respect to time of operation: (1) those running continuously, such as pig iron blast furnaces, non-ferrous metal smelters and refineries, and petroleum refineries; (2) those closing on Sundays and certain important holidays, and operating on the average about 310 days in the year; (3) those closing, in addition, a half day on Saturdays, and operating about 280 days a year.

An aggregative index of volume, it will be recalled, is obtained by multiplying the quantities of the various series in the base period and in the given period by the same set of weights (which weights are prices), summing these base period values (prices × quantities) and these given period values separately, and dividing the latter by the former. As applied to this particular index, the formula is

$$ Q = \frac{\Sigma q_n p_{1923}}{\Sigma q_{1923-1925} p_{1923}}. $$

Shoe production in July 1937 amounted to 34,842,341 pairs. Since there were 23.5 working days in that month in the shoe industry, the average daily amount produced was 34,842,341 ÷ 23.5, or 1,482,653 pairs of shoes. The average daily production for the three years 1923–1925 was 1,167,839 pairs. Since the derived price for shoe production was found to be $1.25,[10] that part of the numerator which refers to shoes is 1,482,653 × $1.25, or $1,853,316; and for the denominator it is 1,167,839 × $1.25, or $1,459,799.

These two value figures, together with others obtained in a similar fashion, and computation of the index number of volume of production for July 1937 are illustrated in Table 154. It should be observed that the index number 113.5 for total leather and products is obtained, not from the two relatives for shoe production and leather tanning, but by relating the two totals for leather shown in this table. In like fashion, the index for total manufactures was constructed from totals for the various manufactured products. Finally, the totals for manufactures and for minerals were combined to produce the index number of industrial production.

In addition to the index described, the Federal Reserve Board makes a second index which is exactly the same, except that seasonal variations

---

[10] The value added by manufacture by the shoe industry was estimated to be $480,000,000 in 1923. Shoe production in that year was 351,114,273 pairs. The price multiplier was therefore $480,000,000 ÷ 351,114,273, or $1.37 for the annual index numbers. However, since the shoe industry operates only 282 days a year instead of 310 days, which was taken as standard, the price multiplier for the monthly index was

$$ \$1.37 \times \frac{282}{310}, \text{ or } \$1.25. $$

TABLE 154

COMPUTATION OF INDEX NUMBERS OF VOLUME OF INDUSTRIAL PRODUCTION, JULY 1937
(Unadjusted index)

Series	$q_{1923-1925}p_{1923}$ (000 omitted)	$q_n p_{1923}$ (000 omitted)	Quantity relative or index number [Col. 3 ÷ Col. 2]
(1)	(2)	(3)	(4)
Leather tanning .. .	$ 939,417	$ 869,967	92.6
All cattle hide leathers.	536,133	472,811	88.2
Calf and kip leathers .	201,347	159,151	79.0
Goat and kid leathers	201,466	238,005	118.1
Shoe production .	1,459,799	1,853,316	127.0
Total leather and products	2,399,216	2,723,283	113.5
TOTAL MANUFACTURES . .	60,639,571	66,991,269	110.4
TOTAL MINERALS . .	10,169,761	11,732,474	115 4
Total industrial production	70,809,332	78,723,743	111.2

Source Division of Research and Statistics, Board of Governors of the Federal Reserve System

are eliminated from the various series before the data are multiplied by the weights. Seasonal variation is eliminated by the moving average method described in Chapter XVII. Changing seasonal is allowed for when appropriate.

Still another feature of these indexes should be mentioned. It was recognized that the relative importance of the different industries in 1923 was far different from that in 1919. Therefore a second index was computed from 1919 to 1922 inclusive, with 1919 weights. This index was then averaged geometrically with the index employing 1923 (or 1923–1925) weights, and this average of the two indexes was considered the final index. For the year 1922, the index with 1923 weights was given a weight twice as great in the average as that with 1919 weights; whereas for 1919, 1920, and 1921, the two indexes were weighted equally. Strictly speaking, this is not an illustration of the "ideal" index number formula, since exact given year weights are not used. But the "ideal" index principle is used, and the index numbers for the years 1919–1922 enjoy the advantages and suffer from the disadvantages of that method.

Federal Reserve Bank of New York Monthly Index of Production and Trade. This index[11] differs in its purpose from that of the Federal Re-

---

[11] This index is not published currently, but mimeographed releases will be mailed upon request to the Federal Reserve Bank of New York.

serve Board Index of Industrial Production in two particulars: (1) It measures the physical volume of *trade* as well as production. It includes not only production (including construction), but everything for which money is paid or for which checks are written, except purely financial activity. The net result is an index with a comparatively small range of fluctuation. Nevertheless, it seems likely that this index exaggerates these fluctuations somewhat, since the series that cannot be obtained (such as personal services) are those that are probably the most stable. (2) It meas‑ures only cyclical movements, excluding secular and periodic movements.

All told, 82 series are used. Before being combined into the finished index, they are adjusted in several ways. As an illustration of the steps involved, the series for mail order house sales has been selected. The numerical illustration running through the following discussion is confined to the year 1937. In order to bring the whole procedure into view at one time, the various operations are summarized in Table 155.

1. *Nearly every series is adjusted to a working day basis; that is, each monthly figure has been divided by the number of working days in that month which are appropriate for the industry in question.* For mail order house sales the following are not considered as working days: Sundays, New Year's Day, Washington's Birthday, Memorial Day, Independence Day, Labor Day, Thanksgiving, and Christmas. If a holiday occurs on Sunday, the holiday is taken on Monday.

2. *Seasonal is allowed for, the seasonal index usually being constructed by the per cent of 12-month-moving-average method.* This is the method used for mail order house sales. Some attention has been given to changing seasonals, also, and in some of the retail series allowance has been made for the variable occurrence of Easter.

3. *Each dollar series is deflated—that is, adjusted for price change.* Since it is difficult to obtain accurate deflating indexes, the use of series requir‑ing adjustment for price changes has been avoided as much as practicable. Mail order house sales, however, necessitated such an adjustment, and a price index was specially constructed for that purpose. Even though the price series may not be so accurate as might be desired, this procedure is better than that of using dollar series entirely uncorrected for price changes. The index with such series included is also more comprehensive than it would be if only physical volume series were used.

4. *Each series is expressed as per cent of normal.* Trend is calculated by whatever method seems appropriate. A formula frequently used is:[12]

$$Y_C = bc^{\frac{1}{d+x}}.$$

---

[12] For the method of fitting this curve, see "A Trend Line for Growth Series, Further Remarks," by Norris O. Johnson. *Journal of the American Statistical Association,* Vol XXXI, December 1936, p. 731.

TABLE 155

DERIVATION OF REFINED SERIES FROM CRUDE DOLLAR FIGURES, MAIL ORDER SALES, 1937

Month (1)	Sales in thousands of dollars (2)	Working days in month (3)	Average sales per working day (4)	Seasonal index (5)	Deseasonalized data (6)	Mail order sales price index (7)	Deflated and deseasonalized data (8)	Trend values (9)	Per cent of normal (10)
January	54,427	25	2,177	71	3,066	77	3,982	3,969	100.3
February	53,831	23	2,340	76	3,079	77	3,999	3,999	100.0
March	78,625	27	2,912	89	3,272	78	4,195	4,029	104.1
April	89,681	26	3,449	99	3,484	78	4,467	4,059	110.1
May	92,627	25	3,705	102	3,632	78	4,656	4,090	113.8
June	89,258	26	3,433	102	3,366	79	4,261	4,120	103.4
July	73,655	26	2,833	86	3,294	79	4,170	4,150	100.5
August	71,254	26	2,741	86	3,187	79	4,034	4,183	96.4
September	90,240	25	3,610	109	3,312	80	4,140	4,217	98.2
October	107,451	26	4,133	122	3,388	80	4,235	4,250	99.6
November	89,813	25	3,593	117	3,071	79	3,887	4,283	90.8
December	116,232	26	4,470	141	3,170	78	4,064	4,316	94.2

Source: Research Department, Federal Reserve Bank of New York

This curve increases at a decreasing percentage rate. The equation for the trend of mail order house sales, fitted to annual data for the years 1922–1932, inclusive, is a straight line fitted to logarithms:

$$\log Y = 3.021189 + .039794X,$$

with origin at 1922, X stated in units of one year, and Y units in thousands of dollars per working day. Of course, it is necessary to extend the trend a few years into the future in order to keep the index up to date.

5. *A few series are smoothed by moving averages.* In the case of construction contracts a 6-month moving average is placed at the sixth month, since current construction is influenced by contracts let any time during the preceding half year, the influence becoming strongest as the current month is reached. No moving average is taken, however, of mail order house sales.

The 82 series, each of which is treated somewhat like mail order house sales, are combined into a general index of trade and production, and into four group indexes as follows:

Production ...	.. 61 series
Primary distribution . . .	9 series
Distribution to consumer .	6 series
Miscellaneous services	6 series
Production and trade.. . .	82 series

The 61 series in the production group index are also combined into the following subgroups, for each of which indexes are constructed.

Producers goods	*Number of series*	*Consumers goods*	*Number of series*	*All goods*	*Number of series*
Durable.	15	Durable	5	Durable.	20
Non-durable. ...	15	Non-durable. ..	25	Non-durable	40
Producers goods	30	Consumers goods	30	Total.	60
Employee hours					1
Production					61

Weights have been derived from Census data and have been based upon total value in trade. The value weight applied to an individual series is often that of a group of activities represented by the series in question, rather than the value in trade of the individual series itself. The weights are not fixed weights but themselves have a secular trend. The weights shown in Table 156, therefore, are those that apply only to the particular period in question. This table shows the method of combining the relatives into index numbers. It should be noticed that the weights used are not exact dollar weights, but only approximate weights in even num-

bers which add up to 100. This facilitates the final combination of the relatives into the finished index numbers. As has been noted, exact weights are not usually of tremendous importance, but they should be approximately correct.

TABLE 156

Construction of Index Number of Production and Trade for December 1937

Series or group (1)	Relative or group index number (2)	Weight of relative or group (3)	Weighted relative or group (4)	Index number (5)
Production				
Durable producers goods .	58	10 9	632 2	
Non-durable producers goods .	78	12 3	959 4	
Durable consumers goods ..	54	5 6	302 4	
Non-durable consumers goods..	91	17.6	1,601 6	
Employee hours	73	8.6	627 8	
Production	...	55.0	4,123 4	75
Primary distribution	...	16 1	1,304 1	81
Distribution to Consumer				
Department store sales	83.8	8 9	745 8	
Chain store grocery sales ..	97.8	5.1	498 8	
Other chain store sales	95.2	3.5	333 2	
Mail order house sales	94.2	2 9	273 2	
Gasoline consumption . ..	97 0	3 2	310.4	
New passenger car registration. .	61.8	2.4	148.3	
Distribution to consumer .	...	26.0	2,309 7	89
Miscellaneous services .	..	2.9	252 3	87
Production and Trade .. .	..	100.0	7,989 5	80

Source. Research Department, Federal Reserve Bank of New York

A major difference between this index and Snyder's Index of the Volume of Trade (of which this is a revision, although it was not made under his supervision) is that, while Snyder's index was intended to cover everything for which money is spent, the present index omits financial activity. There were also other changes in the list of series, as well as changes in weights. Another important source of variation between the two indexes is due to the revision of the trends. The old trend lines, which were fitted before the world depression, were higher during recent years than seemed reasonable. A refitting of the trends to more recent data has had the

effect of lowering the level of the trends for recent years, and therefore of raising the index numbers.[13]

Indexes of business cycles. There are a number of indexes that attempt to measure the cyclical swings of economic activity. The chief differences among these indexes have to do with the data used, the method of computing trend, and the system of weighting. Brief mention will be made of a few of these indexes.

We have seen that the Federal Reserve Bank of New York monthly index eliminates the trend and seasonal from each series separately, and weights the cyclical relatives according to their importance in trade, the weights being revised as the relative importance of the different series gradually changes. None of the indexes below follow exactly this procedure.

Barron's Index of Production and Trade. The technique applied by Warren M. Persons in constructing Barron's Index of Production and Trade is interesting. For the annual index, six series are combined.

(1) The Persons-Day-Thomas index of the physical volume of manufacturing output.

(2) Mineral output.

(3) Building construction.

(4) Electric power production.

(5) Railroad freight traffic.

(6) Wholesale and retail trade.

The index of manufacturing output (described in Census Monograph VIII, 1928, *The Growth of Manufactures*) is so constructed that, for years covered by the Census of Manufactures, it coincides with the Census data.

The different series are expressed as quantity relatives with 1923–1925 as the base. They are then combined into an index by the use of Fisher's "ideal" index number formula. Since the data are relatives, the formula is

$$ Q = \sqrt{ \frac{\Sigma v_{23-25}\left(\dfrac{q_n}{q_{23-25}}\right)}{\Sigma v_{23-25}} \times \frac{\Sigma v_n}{\Sigma v_n\left(\dfrac{q_{23-25}}{q_n}\right)} } \ . $$

The index is thus the geometric mean of two separate indexes: (1) the arithmetic mean of quantity relatives weighted by base year values; (2)

---

[13] For further description of this index, see "New Indexes of Production and Trade," by Norris O. Johnson, *Journal of the American Statistical Association*, Vol 33, June 1938, pp. 341–348. For a description of Snyder's index, see F E. Croxton and D J. Cowden, *Practical Business Statistics*, pp. 390–394, Prentice-Hall, Inc., New York, 1934

the harmonic mean of quantity relatives weighted by given year values. The weights are intended to represent the percentages which the national income from the individual groups (represented by the six series) bears to the total income from all six groups.

A trend is now fitted to the *combined* data. The method, as explained on pp. 411–412, consists of: (1) adjusting the data for population changes; (2) fitting a straight line to these adjusted data; (3) multiplying these straight line values by the estimated population relative to 1923–1925.

A monthly index, using the same series, is computed in a manner similar to that used for the annual index. Seasonal is eliminated for each series separately. The long time trend of the annual index is used for eliminating the trend from the monthly index. A weekly index, based upon less comprehensive data, is also constructed. The data are deseasonalized by dividing each series by its weekly seasonal index. The weights used in combining the series are averages of the base period and given year weights used in the monthly index. The final index is a weighted geometric average of the different series. Both the monthly and weekly indexes are available in two forms since 1919: (1) deseasonalized; (2) adjusted for trend and seasonal.[14]

Index of Business Activity in Buffalo. This is a local index and is published monthly by the University of Buffalo Bureau of Business and Social Research, in its *Statistical Survey*, under the direction of M. A. Brumbaugh. Seven series of local importance are selected. These are adjusted for price changes where appropriate; for variation in calendar, business, or working days; and for secular trend. The trend value for a particular month is the last value of a straight line fitted to the last 19 years for that series. It is thus a moving 19-year straight line, the final rather than the central value of which is taken as the trend value. This method of computing a trend has the advantage of flexibility; and there is not the necessity for occasional revisions which change earlier trend values. On the other hand, less reliance can usually be placed on the end values of a trend fitted by the method of least squares than on the central values, and with the Buffalo index the trend is composed entirely of end values. On the average, the secular trend continues upward until the middle of 1931, after which it turns downward, and flattens out during 1935.

The weighting of each series is directly in proportion to its economic significance, and inversely in proportion to its variability. Economic significance of each series is determined by its representativeness of general business conditions, the amount of employment and buying power it

---

[14] For a more complete description of these indexes, see "Gauging Business Activity," by Warren M Persons, *Barron's*, January 18, 1937, p. 3. Additional references are given in that article.

represents, the value of its product, and other similar factors. As a measure of variability the *average deviation* is used. The reason a series which has great amplitude of fluctuation is reduced in weight is that otherwise such a series might tend to dominate the index.

COMPUTATION OF INDEX NUMBER OF BUSINESS ACTIVITY IN BUFFALO FROM CYCLICAL RELATIVES, JANUARY 1938

Series (1)	Economic significance (2)	Average deviation (3)	Weight [Col 2 ÷ Col 3] (4)	Weight on base of 1.00 (5)	Cyclical relative (6)	Weighted relative (7)
Bank debits	35	12 001	2 916	35	100 0	35 00
Flour milling	5	11 871	421	.05	88 4	4 42
Employment	25	18 765	1 332	16	77 6	12 42
Postal receipts	2	12 388	161	.02	85.1	1 70
New autos registered	6	25.845	232	.03	75 6	2 27
Department store sales	15	6.306	2 379	28	100 6	28 17
Electric power consumption	12	12 050	966	11	92 7	10 20
Total	100		8 437	1 00	.	94 18
Index number			.		.	94 18

Source: University of Buffalo Bureau of Business and Social Research, *Statistical Survey Supplement,* Vol. XIII, No 8A, April 1938, p 6, Table I

The accompanying table illustrates the method of weighting the cyclical relatives. An interesting additional feature of this index is that the different trends are expressed as percentages of 1927 and are averaged together (weighted according to their economic significance), thus producing a composite trend. In showing the index graphically, two lines are shown on one chart: (1) the composite trend index; (2) an index which is the result of multiplying the trend index times the cycle index. The area between the trend line and the trend × cycle line is shaded, so that the resulting silhouette indicates the cycle index.[15]

The idea of allowing the stability of a series to affect its weight is not new. A number of other organizations have been using the idea for some time in constructing indexes of cyclical fluctuations of business. Two such indexes are described in the following paragraphs.

The New York Times Weekly Index of Business Activity. The New York

---

[15] This description is largely a summary of the information appearing in the University of Buffalo Bureau of Business Research, *Statistical Survey, Supplement,* Vol XIII, No. 8A. April 1938.

Times compiles a weekly index composed of seven series.[16]   It assigns to each series an effective weight based upon its relative importance as a business indicator and its reliability, or freedom from erratic or non-business influences.   In order to preserve the effectiveness of these weights, each effective weight is divided by a measure of the cyclical variability of the series.   This measure is the annual average percentage deviation in the cyclical data of the extreme months from the mean of the high and low months.   The series used and their weights are as follows:

Series (1)	Effective weight (2)	Average annual range (3)	[Col. 2 ÷ Col 3] (4)	Adjusted weight (5)
Steel ingot production	25	38	.66	10
Electric power production ..	20	6	3 33	49
Miscellaneous car loadings ..	18	14	1 29	19
Automobile production .	10	56	18	03
Lumber production ..	10	23	.43	.06
Cotton mill activity . . .	10	31	32	.05
Other car loadings . ..	7	13	.54	.08
Total	100	.	6.75	1 00

Each series, before being combined, is adjusted for variation due to working days, for weekly seasonal movements, and for trend.   An interesting feature is that, for most series, secular trend for several years following 1929 was considered inoperative, and a horizontal line used instead.   For some of the series the upward trend has now been resumed.   More specifically, the trends or "normal" values are as follows:

Steel ingot production: 69 per cent of capacity.

Electric power production: Average daily electric power production adjusted for seasonal variation is divided by the adjusted index of steel ingot production with its amplitude reduced to one-fifth.   These monthly figures are smoothed graphically.   The reason for reducing the amplitude of steel ingot production to one-fifth is that the cyclical amplitude of this series is about five times that of electric power production.   This method of obtaining trend is based upon three propositions: (1) that a trend should go approximately through the center of the different cycles; (2) that a proper trend has been discovered for steel ingot production; (3) that the cyclical movements of electric power production and steel ingot production are similar.

---

[16] For a detailed description of this index, see *The New York Times Weekly Index of Business Activity as Revised July 5, 1936*   This pamphlet will be sent upon request to the New York *Times*

Car loadings: The trend of each series (miscellaneous car loadings and other car loadings) was computed in a fashion similar to that for steel ingot production.   In each case this produced a downward trend during the years 1930–1932.

Automobile production: Average daily production for the period 1927–1930.

Lumber production: Average daily production for the period 1929–1931.

Cotton mill activity: Based on percentage of capacity, after a fashion similar to steel ingot production.

The American Telephone and Telegraph Company Index of Industrial Activity.   The American Telephone and Telegraph Company expresses its cyclical deviations in terms of *standard deviations.*   Each series, when so expressed, varies from approximately $+3$ standard deviations to approximately -3 standard deviations.   The series are then averaged together, each being weighted in proportion to its value as a representative of business conditions.   Since the weighted average of all the standard deviations is approximately 10 per cent, each index number is multiplied by 10.   Thus, if the index stands -1.3 standard deviations, it is stated as 13 per cent below normal.[17]

Cleveland Trust Company Index of American Business Activity since 1790. This is the most extensive cycle index, as it extends from 1790 to the present.   Because it was increasingly difficult to find an adequate number of satisfactory series for the earlier years, it was necessary to splice together several sets of annual series over the span covered.

(1)   From 1790 to 1855, 10 series were used.

(2)   From 1855 to 1901, 10 different series were used.

(3)   From 1901 to 1919, the Persons-Day-Thomas index of manufacturing production, with mineral production added, was used.

(4)   From 1919 to date the Federal Reserve index of industrial production was used.

Each series was reduced to a per capita basis and adjusted for trend. Extensive use was made of the high-low mid-point method of computing trend (see pp. 412–418).

The method of obtaining weights is unusual.   Each of the 10 series in the first set was extended through 1882, and each of the 10 series in the second set was extended through 1930.   Then each of the earlier series was compared with the *index numbers* of the latter during the period of overlapping.   The weights assigned to the different series in the earlier

---

[17] For further details, see W. C. Mitchell, *Business Cycles, the Problem and Its Setting,* pp. 295. 328. National Bureau of Economic Research, New York, 1927.

period were in proportion to their closeness of correspondence to the movements of the index numbers. (Technically, the weights assigned were in proportion to the coefficients of correlation which were computed. See Chapter XXII for an explanation of correlation.) Before the different series were averaged together, they were put in terms of their average deviations.

Greater difficulty still was occasioned in obtaining monthly data; from 1790 to 1815, for instance, it was necessary to rely entirely on commodity prices. The different monthly series selected were fitted to the annual index numbers so as to make the average monthly values of the former coincide with the annual index numbers. Large silhouette charts (such as Chart 24), as well as the monthly index numbers, which are deviations from normal, are published by the Cleveland Trust Company, of Cleveland Ohio, and are available upon request. A description of the index by its author, Col. Leonard P. Ayres, is also on the large chart.

Indexes of Qualitative Changes or Differences

Adequacy of State Care of Mental Patients. Such an index comparing the different states was constructed for the years 1922 and 1933 by Ellen Winston and published in the *American Sociological Review* of April 1938.[18]

The index is a simple average of five sets of relatives:

(1) Nurses and attendants per 1,000 average daily resident patient population (125 nurses and attendants per 1,000 patients = 100).

(2) Physicians per average daily resident patient population (6.67 physicians per 1,000 patients = 100).

(3) Physicians per annual admissions (25 physicians per 1,000 admissions = 100).

(4) Annual cost of maintenance per average daily resident patient population ($312 = 100).

(5) Value of hospital property per average daily resident patient population ($1,500 = 100).

Two group indexes were also computed, the first three series being averaged together to obtain an index of personnel in state hospitals, and the last two to obtain an index of expenditures of state hospitals. These two group indexes corresponded closely with each other.

Since the same standard or base was used for 1922 and 1933, it is possible not only to compare the different states for the same year, but also to compare the adequacy of a given state's care in the two years. No index number, however, is computed for the United States as a whole.

---

[18] "Indices of Adequacy of State Care of Mental Patients," by Ellen Winstor *American Sociological Review,* Vol 3, April 1938, pp. 190–202.

The fairness of the comparisons between states and between periods of time is somewhat impaired by the fact that no adjustment is made in the two financial series for variations in the value of the dollar between states or between years.

Measures of adequacy of state school systems. All but one of the preceding illustrations have dealt with measurements of primary interest to economists, and for the most part have dealt with comparisons over a period of time. This concluding section on index numbers will have to do with school systems, and will describe some index numbers which compare the adequacy of such systems in the different states. Although no very satisfactory measure has yet been devised, a number have been undertaken, and the following pages will illustrate some interesting variations in procedure.

Ayres' Index of State School Systems. Probably the first comprehensive index of school systems was undertaken by Leonard P Ayres in 1912. A revision was made in 1920.[19] In the revised index ten series of data were averaged together for each state, five of the items having to do with attendance and five with financial matters. These items and the multipliers used to reduce them to a comparable basis are:

	Measure of Adequacy	*Multiplier*
1	Per cent of school population attending public schools daily	1
2	Average number of days attended by each child of school age	$\frac{1}{2}$
3	Average number of days public schools were kept open	$\frac{1}{2}$
4.	Per cent that high school enrollment is of total enrollment (2.75 multiplier in states with 11-year systems)	3
5.	Per cent that boys are of girls in high schools	1
6.	Average annual expenditure per child attending	1
7.	Average annual expenditure for each child of school age	1
8.	Average annual expenditure per teacher employed	$\frac{1}{2\frac{1}{4}}$
9.	Expenditure per pupil for purposes other than salaries	2
10.	Expenditure per teacher for salaries	$\frac{1}{1\frac{1}{2}}$

The object of these multiplications or divisions was to express each series for each state as a per cent of some standard which was considered desirable. Thus 200 days was considered the standard length of school year; consequently if a school in a particular state kept open 200 days, that state would have a rating of 100 with respect to item 3. It was possible, and it occasionally happened, that a state exceeded the standard set for a particular item and therefore rated higher than 100. The index numbers were simple arithmetic averages of these relatives. Index numbers were

---

[19] See Russell Sage Foundation, Circular No 124, *A Comparative Study of Public School Education in the 48 States;* and Leonard P. Ayres, *An Index Number for State School Systems,* Russell Sage Foundation, New York, 1920.

computed for the United States as a whole for each year from 1871 through 1918, and for each state for the years 1890, 1900, 1910, 1916, and 1918. Since the index has been brought up to date by Frank M. Phillips, index numbers for each state are now available for the additional years 1920 and 1922.

In addition to bringing the Ayres index up to date, Phillips made a revision employing an additional technique for the years 1910, 1918, 1920, 1922, 1924, and 1930. This consisted in adjusting the five financial series for changes in cost of living. Each series was deflated by dividing by the United States Bureau of Labor Statistics Cost of Living Index (for 1910, the Retail Food Price Index was used). This adjustment was for the purpose of making the index numbers of state school systems more comparable over time, so that an increase in money expenditures would produce an increase in the index number only to the extent that the former meant an increase in real expenditures also. The effect of the adjustment has been that, both for the United States as a whole and for the different states, the school systems have shown considerable less improvement over time. Thus for the United States as a whole we have the following results:

Year		Original Index	Revised Index
1910	...	42.41	43 55
1918	. ...	51 01	44 34
1920	....	59.42	44.73
1922	...	74.50	57 15

The revised index is designed to show two-way comparisons: among states, and over time. The comparisons among states are not completely valid, however. Although the financial items have been adjusted for changes in the cost of living over time, they have not been adjusted for differences in cost of living among states. It is also true that costs of living vary directly with density of population and degree of urbanization Phillips took cognizance of the density and urbanization factors by publishing a supplementary table in which the states are classified into groups on these bases and the rank of the different states in each group is shown.[20]

Phillips' Index of Educational Rank   Phillips also computed a new index based upon the following items:

(1)   Percentage of illiterates in the population ten years of age and over.

(2)   Ratio of the number of children in average daily attendance in public schools to the number of age 5 to 17 inclusive.

(3)   Per cent that high school enrollment is of total enrollment.

---

[20] See Frank M. Phillips, *Educational Ranking of States by Two Methods*, Bruce Publishing Company, Milwaukee, 1925; and "Educational Rank of States, 1930," *The American School Board Journal*, February, March, and April, 1932, Vol. 84, Nos. 2, 3, 4.

(4)   Average number of days attended by each child enrolled.

(5)   Average number of days schools were kept open.

(6)   Ratio of the number of students taking teacher-training courses to the number of teaching positions.   (For the 1930 index a different series was substituted.)

(7)   Per cent of high school graduates continuing their education.

(8)   Total cost, excluding salaries, per pupil in average daily attendance.

(9)   Average annual salary of teachers, principals, and supervisors

(10)   Total amount expended per child of school age.

It is seen that Phillips retains a number of Ayres' series.   He adds a series on illiteracy, one on teacher training, and one on higher education. On the other hand, he reduces the number of financial series to three. These are not adjusted for changes in the cost of living, since the Phillips index is designed to make comparisons among states for a given year only. The method of constructing the index is different also.   The states are first ranked separately with respect to each given criterion; then the ranks are summed, state by state; finally the states are ranked according to the sum of their ranks.   For instance, in 1930 Washington ranked as follows with respect to the ten different criteria: 3, 4, 1, 28, 13, 7, 20, 18, 14, 11. These ranks total 119.   Since this was the lowest total, Washington was first in educational rank among the states.

Item 6, the ratio of teacher-training students to teaching positions, was discarded in 1930, since a large ratio may in practice sometimes mean that the teaching field is being overcrowded by certifying poorly qualified teachers.   In place of this item there was substituted, where obtainable, the per cent of teachers employed who had at least two years' training beyond high school graduation; where these data could not be obtained, the ratio of teachers to students was used.   Item 7, the per cent of high school graduates continuing their education, is not very satisfactory either. In general, a state that ranks high with respect to this criterion ranks low with respect to the others.   In technical language, there is negative correlation between this criterion and *each* of the others.   (See Chapter XXII.) The illiteracy series is still another which is not considered appropriate by some authorities.

Rankings are available for the sum of ranks for 1910, 1918, 1920, 1922, 1924, and 1930, and for each criterion for most of these years.[21]

N.E.A. Ranking. ·The Research Division of the National Education Association has selected five criteria for judging the adequacy of school systems.   Only those series have been selected that are widely accepted

---

[21] *Ibid.;* also "Educational Ranking of States by Two Methods," by Frank M Phillips, *American School Board Journal,* Vol. 69, December 1924, pp. 47–49.

as valid, whose validity is supported by the data available, that are reasonably reliable, and for which comparable nation-wide figures are available. No attempt has been made to combine the different series into an index, since information is not available for their proper weighting. The five series selected were:

(1) The proportion of children reached by the services of the schools, measured by the ratio of actual number of student days in school to the standard number of student days considered desirable

(2) The holding power of schools, measured by the ratio of children aged 14–17 (high school age) attending school to the number of children aged 14–17.

(3) The quality of teaching provided, measured by the average salary paid teachers, principals, and supervisors.

(4) The material school environment, measured by the value of school property per child enrolled.

(5) The per cent of literacy among native-born population over 10 years of age.

The states are ranked for the year 1930 with respect to each of these criteria; the table of rankings was published in the *National Education Association Research Bulletin*, May 1932 (p. 126)

N.E.A. Index of Financial Adequacy. Confronted with the question of whether federal aid to states for education was desirable, the Research Division of the National Education Association undertook to construct measures of effort expended by states for education and the adequacy of the results obtained. The Association abandoned any attempt to measure adequacy by combining separate measures of the type we have been describing, and substituted instead the following ratio:

$$\frac{\text{Amount spent for education}}{\text{Units of educational need}}.$$

The numerator of this fraction is: total expenditures for current expenses, exclusive of interest (which is not available for all states), plus cost of state department of education, less amount of expenditures received from the Federal Government and from subsidies. The denominator is Mort's Index of Educational Need.[22]

The method followed by Mort in constructing his index is very ingenious. The unit of educational need is one student attending elementary schools daily. It is recognized, however, that it is more expensive to support a school system in a rural community than it is in an urban com-

---

[22] "An Objective Basis for the Distribution of Federal Support to Public Education," by Paul R. Mort, Teachers College Record, Vol. XXXVI, November 1934, pp. 91–110.

munity, and more expensive to support the education of a high school student than that of an elementary school student.   Consequently certain adjustments must be made.   First, the educational need attributable to a high school student is considered 1.7 times that of an elementary student. The adjustment for additional need attributable to residence in a rural territory is more complicated.   A community is considered rural territory if it has 2,500 population or less   As a measure of degree of rurality for a given state, the ratio of number of acres of farm land per inhabitant 5 to 20 years of age living in rural territory is computed.   Let us call this the *rurality ratio*.   The correction factor for additional need due to rurality is obtained by taking the sum of 1.22 plus four-thousandths times the rurality ratio (provided, however, that the correction factor shall not exceed 1.70). This may be expressed as the following equation, in which Y_C is the correction factor, for a state, and X is the rurality ratio:

$$Y_C = 1.22 + .004X,$$

with the maximum Y_C value of 1.70.   It is evident that a high school student in daily attendance in a rural territory in a state with the maximum rurality ratio constitutes an educational need of $1.7 \times 1.70 = 2.89$, compared with a value of 1.00 for an elementary student in an urban territory in a state with a minimum rurality ratio.

An additional correction is made in the Mort Index of Educational Need for variation in cost of living as between communities of different size. The correction for cost of living varies from 30 per cent for communities of more than 500,000 population to no correction for communities of less than 10,000 population.   This correction is not entirely satisfactory since it does not allow for variation among different sections of the United States.   The state index numbers of educational need are deflated in the usual fashion by the cost of living index.

The index of financial adequacy obtained by dividing the amount spent for education by Mort's index is especially suitable for the purpose for which the index is intended.   If it can be shown that the financial adequacy of a state system bears little relationship to the amount of financial effort or sacrifice which a state makes, that constitutes a good talking point in favor of Federal aid to states for education.   Effort is defined as "the extent to which a state extends itself to support education in terms of its financial ability," and is measured by the ratio:

$$\frac{\text{Amount spent for education}}{\text{Financial resources}}.$$

The measurement of financial resources itself constitutes an interesting problem in index number construction, but will not be discussed in this

volume. It is the conclusion of the National Education Association that these measures constitute a valid argument for federal aid.[23]

Sources of Current Index Numbers

A considerable proportion of the most useful economic indexes that are published currently can be found in one or more of the following periodicals:

(1) *Survey of Current Business*, published by the United States Department of Commerce, Bureau of Foreign and Domestic Commerce. Back data are available in the different annual supplements.

(2) *Federal Reserve Bulletin*, published by the Board of Governors of the Federal Reserve System.

(3) *Standard Trade and Securities Basic Statistics*, Volume 3, Statistical Section, and *Current Statistics;* published by Standard Statistics Company, Inc.

Other sources of statistical data, many of which publish index numbers, will be found in Appendix A. For a comprehensive list of indexes, together with brief description and sources, the reader is referred to Donald H. Davenport and Frances V. Scott, *An Index to Business Indexes*, Business Publications, Inc., Chicago, 1937.

Selected References

W. L. Crum, A. C. Patton, and A. R. Tebbutt. *Introduction to Economic Statistics,* Chapter XIX; McGraw-Hill Book Co , New York, 1938.

H. T. Davis and W. F. C. Nelson: *Elements of Statistics,* Chapter IV, Principia Press, Bloomington, Indiana, 1935.

Irving Fisher· *The Making of Index Numbers,* Houghton Mifflin Co., Boston, 1927 A clear and authoritative exposition of one school of thought

J. M. Keynes: *A Treatise on Money,* Volume I, Chapter VIII; Macmillan Co., New York, 1930.

Willford I. King· *Index Numbers Elucidated;* Longmans Green, New York, 1930 King disputes Fisher's contention that an index number formula can be selected on the basis of certain mathematical tests. Rather, he says, its purpose dictates the formula to use.

F. C. Mills: *Statistical Methods Applied to Economics and Business* (Revised Edition), pages 199–224 and Chapter IX; Henry Holt and Co., New York, 1938.

W. M. Persons: *The Construction of Index Numbers,* Houghton Mifflin Co., Boston, 1928. A concise discussion of the relative advantages of different formulae. In general, the argument supports Fisher.

C. M. Walsh· *The Measurement of General Exchange Value;* Macmillan Co., New York, 1901. A pioneer study of index number methodology.

---

[23] See "The Efforts of the States to Support Education," in *National Education Association Research Bulletin*, Vol. XIV, May 1936, pp. 103–163. The issue was prepared by Lyle W Ashby.

CHAPTER XXII

SIMPLE CORRELATION

One of the chief objectives of science is to estimate values of one factor by reference to the values of an associated factor. "The scientific method . . . consists in the careful and laborious classification of facts, in the comparison of their relationship and sequences, and finally in the discovery by the aid of disciplined imagination of a brief statement or *formula*, which in a few words resumes a wide range of facts. Such a formula . . . is termed a scientific law."[1] When the relationship is of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as *correlation*.

A Simple Explanation

It may surprise some of us to know that there is a very close relationship between temperature and the frequency with which crickets chirp. If, for instance, we should count the number of chirps made by a cricket in 15 seconds and add it to 37, we could closely approximate the Fahrenheit temperature at that time. Or, if we should multiply the degrees Fahrenheit by 3.78 and subtract 137 from the result, we could estimate the number of chirps to be expected from a cricket in one minute. This relationship would be found remarkably accurate, unless the temperature was below 45°. When the weather is colder than 45°, crickets do not chirp. Likewise, it might not be accurate appreciably beyond 80° since observations have not been made beyond that temperature and we do not know, therefore, if the relationship holds for higher temperatures.

The relationship between these two variables—temperature and cricket chirps—is displayed in Chart 215, known as a *scatter diagram*. Each dot represents an observation of one cricket. Thus observation A represents a cricket which, at a temperature of 59.0°, chirped 85 times per minute. The reader should notice that temperature is plotted along the X-axis,

---

[1] Karl Pearson, *The Grammar of Science*, p 77. Adam and Charles Black, London, 1900.

while chirps per minute are plotted along the Y-axis. This is because the number of chirps per minute appears to be a direct result of the temperature. In this case it is also true that we wish to estimate the number of chirps to be expected at a given temperature. Temperature is therefore the independent variable, and chirps per minute the dependent variable. Even though it were temperature we wished to estimate, it would nevertheless be best to show the causal factor on the X-axis. When the causal relationship is not clear or when neither factor can be said to be the cause
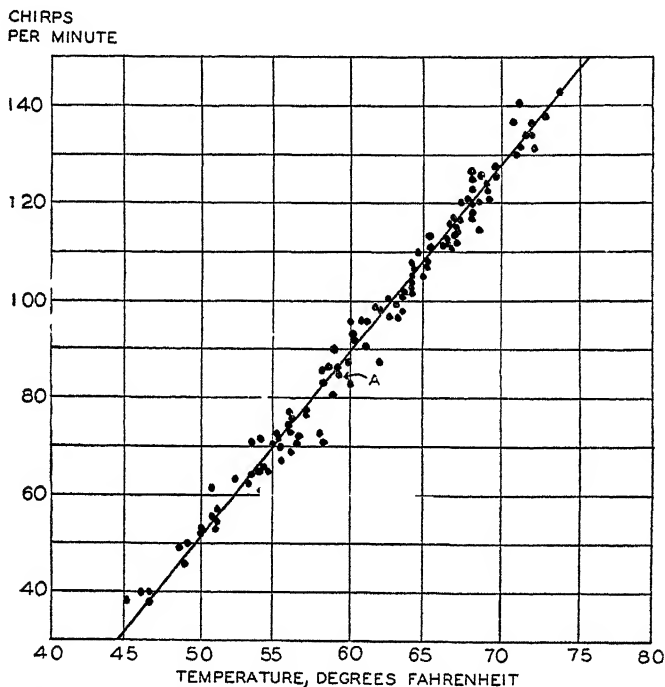
CHIRPS
PER MINUTE

Chart 215. Temperature and Chirps per Minute of 115 Crickets. (Data provided by Mr. Bert. E. Holmes.)

of the other, then the variable to be predicted should be plotted on the Y-axis.

Judging from Chart 215, we see that the relationship between the two variables is linear, for the straight line appears to be as good a fit as a more complicated curve. The equation of this line[2] is

$$Y_C = -137.22 + 3.777X.$$

---

[2] This equation was fitted by the authors to data furnished by Bert E. Holmes. See also Bert E. Holmes, "Vocal Thermometers," *The Scientific Monthly*, Vol. XXV, September 1927, pp. 261–264.

From this equation, estimates of chirps can be made for any desired temperature within the limits of the observations shown on the chart. Thus, if we wish to estimate the number of chirps when the temperature is 59.0° (observation A), we find the number by substituting 59.0 for X in the equation. Thus

$$Y_C = -137.22 + (3.777)(59.0) = 86 \text{ chirps.}$$

The estimate could be read, although less accurately, directly from the estimating line plotted on the chart. Although the estimate (86) does not agree perfectly with the actual observation of 85 chirps, the discrepancy is not large.

We cannot fail to be impressed with the adequacy of the generalization expressed in the equation $Y_C = -137.22 + 3.777X$. Since most of the dots are very close to the line, it appears that frequency of chirps has been adequately explained by reference to temperature. The slight variations from the estimating line are unexplained and may be due to differences between individual crickets, differences associated with the time of day or year in which the observations were made, humidity, and inaccuraries of observation of temperature or number of chirps. Also, the temperature at the spot where the cricket is chirping may be different from that where the observer is standing. This might be the case if the cricket were under a stone. An examination of other causes of variation, in addition to temperature, involves consideration of three or more variables, a procedure for which will be considered in Chapter XXIV under the heading of multiple correlation.

The closeness of the relationship may be expressed in general terms by stating that the *coefficient of correlation*, r, is $+.9919$. Since ± 1.0 is perfect correlation and 0 is no correlation, we can readily imagine that one almost never finds a higher coefficient than $+.9919$. The plus sign indicates that the correlation is positive—that is, that the chirps increase as the temperature increases. Had chirps decreased with increasing temperature, the correlation would have been negative, or inverse; the sign of r would have been negative, as would the sign of b in the estimating equation; and the estimating line would have sloped downward to the right.

An illustration of rather low correlation $(-.11)$ is given by Chart 216. In this case, brain weight was estimated by cranial capacity, and legislative ability by a rather complicated system of scoring. But even if we assume that all measurements are accurate, the evidence certainly does not suggest that legislators should be selected solely from head measurements. Perhaps there are additional factors which account for legislative ability; for example, intelligence, education, initiative, honesty, social awareness, and other traits are doubtless important.

Correlation Theory

Correlation may be thought of as involving three types of measurements, which may conveniently be made in the following order:

(1)   An *estimating equation* which describes the functional relationship between the two variables.   As the name indicates, one object of such an equation is to make estimates of one variable from another.
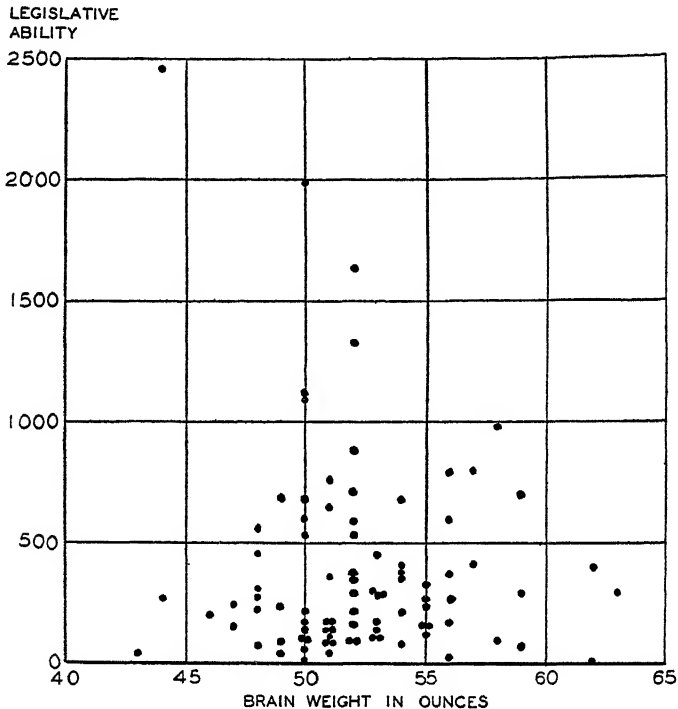
Chart 216.   Estimates of Brain Weight and Legislative Ability of 89 Members of Congress.   (Data from "Brain Weight and Legislative Ability in Congress," by Arthur MacDonald, *Congressional Record*, April 12, 1932)

(2)   A measure of the amount of variation of the actual values of the dependent variable from their estimated or computed values.   This measure of the variation which has not been explained by the estimating equation is analogous to a standard deviation and gives an idea, in *absolute* terms, of the *dependability of estimates*.   It is called the *scatter*, or *standard error of estimate* (σ_{y_s}).

(3)   A measure of the *degree* of relationship, or *correlation* (r), between the variables, independent of the units or terms in which they were originally expressed.   A closely related measure (r^2) will permit us to state

the *relative* amount of variation which has been explained by the estimating equation.

The estimating equation.   Foresters sometimes find it convenient to estimate the height growth of trees from their growth in diameter, since this procedure is quicker than direct measurements of the growth in height. The scatter diagram, Chart 217, shows the breast-high diameter growth and the growth in height of 20 trees, together with the estimating line which describes the nature of the relationship between the two variables. This straight line has been so fitted that the sum of the squares of the Y deviations from it is less than those from any other straight line.   A
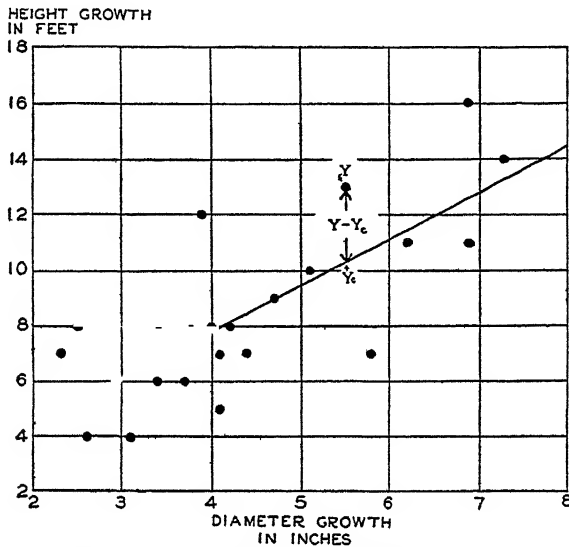
Chart 217.   Breast-High Diameter Growth and Height Growth of 20 Forest Trees.
(Data of Table 157.)

curve fitted in this manner is usually considered by statisticians to be the best with which to estimate values of one variable when values of the other variable are known.   The fitting of such a line is similar to the fitting of a trend and requires the use of the following normal equations:[3]

$$\text{I.} \quad \Sigma Y = Na + b\Sigma X.$$
$$\text{II.} \quad \Sigma XY = a\Sigma X + b\Sigma X^2.$$

Table 157 shows the computations that are necessary to determine the values which must be substituted.   The substitution yields.

$$\text{I.} \quad 173 = 20a + 90.7b.$$
$$\text{II.} \quad 856.0 = 90.7a + 453.93b.$$

---

[3] The normal equations were discussed in Chapter XV.

Multiplication of all the items in equation I by 4.535 permits us to cancel out a by subtracting equation I from equation II.　Thus

$$\text{I} \times 4.535. \quad 784.555 = 90.7a + 411.3245b$$
$$\text{II.} \quad 856\ 0 \quad = 90.7a + 453.93b$$
$$\overline{71.445 = \qquad\qquad 42.6055b}$$
$$b = 1.677.$$

We may now substitute the value of b in equation I in order to find the value of a.

$$\text{I.} \quad 173 = 20a + 152.1039$$
$$a = 1.045.$$

TABLE 157

COMPUTATION OF VALUES USED IN COMPUTING ESTIMATING EQUATION FOR GROWTH IN DIAMETER AND HEIGHT OF 20 FOREST TREES

Rank in diameter growth (smallest to largest)	Diameter growth at breast height in inches X	Height growth in feet Y	XY	X^2
1	2.3	7	16 1	5 29
2	2.5	8	20.0	6.25
3	2 6	4	10 4	6.76
4	3.1	4	12 4	9.61
5	3.4	6	20 4	11 56
6	3 7	6	22.2	13 69
7	3 9	12	46 8	15.21
8	4.0	8	32 0	16 00
9	4.1	5	20 5	16 81
10	4.1	7	28.7	16 81
11	4.2	8	33 6	17.64
12	4.4	7	30 8	19 36
13	4 7	9	42.3	22.09
14	5 1	10	51 0	26.01
15	5.5	13	71 5	30.25
16	5.8	7	40 6	33.64
17	6.2	11	68 2	38 44
18	6.9	11	75.9	47 61
19	6.9	16	110 4	47.61
20	7.3	14	102 2	53.29
Total	90.7	173	856 0	453.93

Source　Donald Bruce and F　X. Schumacher, *Forest Mensuration*, p　124, McGraw-Hill, New York, 1935.　Courtesy of Publisher and Authors.

The values for a and b are checked by substitution in equation II.　While this does not prove that no errors in computation have been made, yet if the correct numbers were substituted in the two normal equations, either

no errors, or counterbalancing errors, have been made. Since $a = 1.045$ and $b = 1.677$, the equation of the line which enables us to estimate the growth in height of trees in this particular forest when their growth in diameter is known may be stated as

$$Y_C = 1.045 + 1.677X.$$

Suppose now we wish to estimate the height growth of a tree which grew 5.5 inches in diameter. Substituting in the equation, we have

$$Y_C = 1.045 + (1.677)(5.5)$$
$$= 10.268 \text{ feet.}$$

Dependability of estimates. However, we should not expect all trees which grew 5.5 inches in diameter to have grown exactly 10.268 feet in height, for the dots of the scatter diagram do not all lie on the fitted line. Rather, 10.268 should be thought of as an estimate of the average height growth of all trees of the diameter growth indicated. We should expect variations from this value the same as from the arithmetic mean of a frequency distribution. It is therefore pertinent to inquire what proportion of trees may be expected to fall within any range of error in which we may be interested, assuming, of course, that we have a representative sample.

To do this, it is necessary to compute the standard deviation of the Y values, not from their mean, but from the line of estimation. On Chart 217 the vertical distance from the line of estimate to any Y value represents the difference between the observed Y value and the estimated Y value. The estimated Y values, Y_C, are obtained by solving the estimating equation for each measurement of diameter growth, or X value. The deviation $Y - Y_C$ represents the error that would have been made in one particular instance. To obtain a summary measure of those deviations, they may be squared, summed, and divided by N, and the square root extracted. This is the *scatter*, or *standard error of estimate* and may be denoted[4] by σ_{y_s} or S_Y. Its formula may be written

$$\sigma_{y_s} = \sqrt{\frac{\Sigma(Y - Y_C)^2}{N}}.$$

In this illustration $\sigma_{y_s} = \sqrt{\dfrac{88.75}{20}} = \sqrt{4.44} = 2.107$. Calculations are

---

[4] The symbol S_Y is frequently used for this concept. Although this measure is frequently spoken of as a "standard error of estimate," it is not a standard error in the sense used in Chapters XII and XIII. σ_{y_s} is the standard error of an individual item when the values are measured as deviations from the estimated values (Y_C) in the same sense that σ_y is the standard error of an individual item when the values are measured as deviations from their mean ($\bar{Y}$). Consequently it seems more logical to use the symbol σ_{y_s} than S_Y.

shown in Table 158, columns 7 and 10. Ordinarily the more expeditious method of calculation which is explained on page 671 would be used. The above method is used solely to explain the meaning of the measure.

This measure may be interpreted in a manner strictly analogous to that of the standard deviation of a frequency distribution. Although σ_{y_s} is called the standard *error* of estimate, it is not to be thought of as a measure of the variation of estimates made from different samples, but merely as a general measure of the variation of the actual Y values from the computed Y values of a particular sample. It yields an estimate of the range above and below the line of estimation within which 68.27 per cent of the
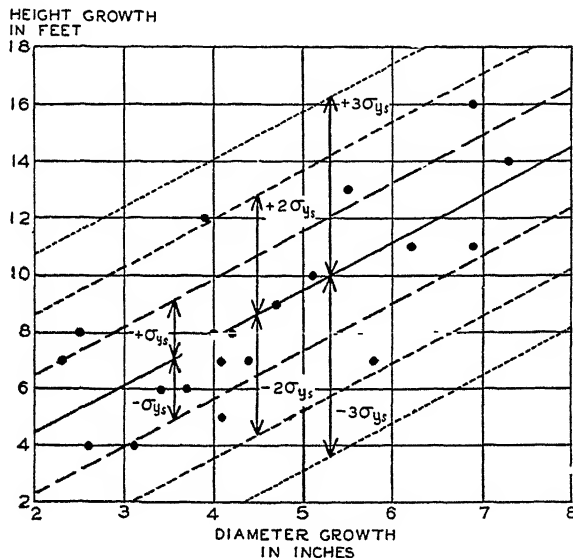
Chart 218.   Estimating Line and Zones of Scatter for Diameter Growth and Height Growth of 20 Forest Trees.   (Data of Table 158.)

items may be expected to fall if the scatter is normal. In practice we frequently think of this measure as the range within which about $\frac{2}{3}$ of the values will be found. For the case in hand ($\sigma_{y_s} = 2.11$), we may expect to find about $\frac{2}{3}$ of the items of Chart 218 within the narrow band $\pm\sigma_{y_s}$ shown in the diagram; about 95 per cent (ideally 95.45) within the wider band that includes $\pm 2\sigma_{y_s}$; and practically all within $\pm 3\sigma_{y_s}$ (theoretically, with a large number of items, 99.73 per cent of the cases). A count of the dots shows that within $\pm\sigma_{y_s}$ of the line of estimate, 13 of the 20 items (65 per cent) are found; within $\pm 2\sigma_{y_s}$ of the line, 19 of the items (95 per cent) appear; and within $\pm 3\sigma_{y_s}$ are included all 20 of the items. The slight discrepancies may have been due to the fact that the sample was small

TABLE 158

COMPUTATION OF TOTAL VARIANCE, EXPLAINED VARIANCE, AND UNEXPLAINED VARIANCE, FOR HEIGHT GROWTH OF 20 FOREST TREES AS ESTIMATED BY THEIR DIAMETER GROWTH

Rank in diameter growth (smallest to largest) (1)	Diameter growth at breast height in inches X (2)	Height growth in feet Y (3)	Y_c (4)	Deviations			Squared deviations		
				$y = Y - \bar{Y}$ (5)	$y_c = Y_c - \bar{Y}$ (6)	$y_s = Y - Y_c$ (7)	$y^2 = (Y - \bar{Y})^2$ (8)	$y_c^2 = (Y_c - \bar{Y})^2$ (9)	$y_s^2 = (Y - Y_c)^2$ (10)
1	2 3	7	4.902	−1 65	−3 748	2 098	2 7225	14 0475	4 4016
2	2 5	8	5.238	−0 65	−3 412	2 762	0 4225	11 6417	7 6286
3	2 6	4	5.405	−4 65	−3 245	−1 405	21 6225	10 5300	1 9740
4	3 1	4	6.244	−4 65	−2 406	−2 244	21 6225	5 7888	5 0355
5	3 4	6	6.747	−2 65	−1 903	−0 747	7 0225	3 6214	0 5580
6	3 7	6	7.250	−2 65	−1 400	−1 250	7 0225	1 9600	1 5625
7	3 9	12	7.585	3 35	−1 065	4 415	11 2225	1 1342	19 1922
8	4 0	8	7.753	−0 65	−0 897	0 247	0 4225	0 8046	0 0610
9	4 1	5	7.921	−3 65	−0 729	−2 921	13 3225	0 5314	8 7322
10	4 1	7	7.921	−1 65	−0 729	−0 921	2 7225	0 5314	0 8182
11	4 2	8	8.088	−0 65	−0 562	−0 088	0 4225	0 3147	0 0077
12	4 4	7	8.424	−1 65	−0 226	−1 424	2 7225	0 0511	2 0278
13	4 7	9	8.927	0 35	0 277	0 073	0 1225	0 0767	0 0053
14	5 1	10	9.598	1 35	0 948	0 402	1 8225	0 8987	0 1616
15	5 5	13	10.268	4 35	1 618	2 732	18 9225	2 6179	7 4638
16	5 8	7	10.772	−1 65	2 122	−3 772	2 7225	4 5029	14 2280
17	6 2	11	11.442	2 35	2 792	−0 442	5 5225	7 7953	0 1954
18	6 9	11	12.616	2 35	3 966	−1 616	5 5225	15 7292	2 6115
19	6 9	16	12.616	7 35	3 966	3 384	54 0225	15 7292	11 4515
20	7 3	14	13.287	5 35	4 637	0 713	28 6225	21 5018	0 5084
Total	90 7	173	173 001	0	0 004	−0 004	208 5500	119 8085	88 7548
Mean	4 535	8 650	8 650	..	0 004	.	10 43	5 99	4 44
σ	...	...	.	..			3 229	2 448	2 107

Source: Data of Table 157.

and the scatter not normally distributed around the estimating equation.

It was calculated that trees with growth in diameter of 5.5 inches should average 10.268 feet in height growth. We may now amplify the statement by saying that, if our sample is representative, about $\frac{2}{3}$ of such trees should vary in height growth between 8.16 feet and 12.38 feet (10.268 ± 2.107); or, considering a slightly wider range, about 95 out of 100 should lie between 6.05 feet and 14.48 feet. The proportion lying within any other range could readily be computed also by referring to Appendix E.

These statements concerning range of error have to do, not with certainty, but only with expectation. We have used only 20 items, and, even though the sample may have been carefully chosen, another sample of 20 would not give us precisely the same results as those obtained above. It might be that we could reduce uncertainty further, not only by increasing the size of our sample but also by comparing variations in height growth with some other factor in addition to diameter growth—for example, age, since as trees grow older their rate of growth may change. Also, the character and quantity of plant food in the soil and the degree of crowding of the trees might be considered. Even if several factors in addition to diameter growth were considered, there would still be some unexplained variations, and therefore still some uncertainty.

The correlation coefficient and explained variability. Another measure closely related to the estimating equation and scatter, and frequently used in the social sciences is the coefficient of correlation r. The estimating equation $Y_c = a + bX$ is a statement of the way in which the dependent variable changes with variations in the independent variable. σ_{y_s} is an indication of the amount of dispersion in the dependent variable which we have failed to account for by our line of estimation, but it is stated in terms of the original data—in the case of the diameter growth and height growth data, in feet. When stating the degree of relationship between two variables, it is convenient to be able to state results in concise numerical terms which are independent of the units of the original data, and to express the degree of relationship between two series even though we do not know the equation of the line of estimation or σ_{y_s}. To be sure, something is lost by so compressing the information, since it does not enable us to make an estimate of the value of one variable from the other, or to tell, in absolute magnitude, the degree of accuracy of any prediction we may make. But something is gained too, since one coefficient can be compared with any other, regardless of the subject matter of the different correlations. As has been stated, the coefficient of correlation is a number varying from +1, through zero, to −1. The sign indicates whether the slope of the line of relationship is positive or negative, while the magnitude

of the coefficient indicates the degree of association. When there is absolutely no relationship between the variables, r is 0.

A clear understanding of the meaning of the coefficient of correlation is given by the following approach. One measure of variability, called *variance* or *total variance*, is the square of the standard deviation of the Y values. This total variance can be broken up into two parts: that which has been explained by our line of relationship, and that which we have failed to explain. The *total variance* in height growth of the trees of our distribution, as indicated by the calculations in column 8 of Table 158, is $(3.229)^2$, or 10.43. The amount of variability which we have explained by our line of relationship may likewise be measured by the square of another standard deviation, that of the estimated Y values from their own mean (which is also the mean of the original Y values)[5]. The *explained variance* is shown in column 9 of Table 158 to be $(2.448)^2 = 5.99$. The *unexplained variance* is the square of the standard error of estimate. But this measure has already been found to be 2.107; hence the unexplained variance is $(2.107)^2 = 4.44$.

Let us summarize our findings:

Variance	Symbol and formula		Amount of variance	Per cent of total variance
Unexplained....	$\sigma_{y_s}^2 = \dfrac{\Sigma y_s^2}{N} =$	$\dfrac{\Sigma(Y - Y_c)^2}{N}$	4 44	42.6
Explained.........	$\sigma_{y_c}^2 = \dfrac{\Sigma y_c^2}{N} =$	$\dfrac{\Sigma(Y_c - \overline{Y})^2}{N}$	5.99	57.4
Total	$\sigma_y^2 = \dfrac{\Sigma y^2}{N} =$	$\dfrac{\Sigma(Y - \overline{Y})^2}{N}$	10.43	100 0

It may be helpful to some readers also to visualize this information Chart 219 shows for the data of height growth:

A.   The derivation of total variance, which is based upon the deviations of the actual Y values from their mean.

B.   The derivation of explained variance, which is based upon the deviations of the computed Y values from their mean. (Note that $\overline{Y}_c = \overline{Y}$.)

C.   The derivation of unexplained variance, which is based upon the deviations of the actual Y values from the computed Y values.

The values of each standard deviation and variance are shown at the right. The variances are indicated by shaded squares, and it is to be observed that the sum of the areas of the two lower squares equals that

[5] See Appendix B, section XXII–1, equation 2.
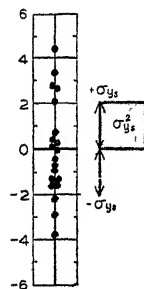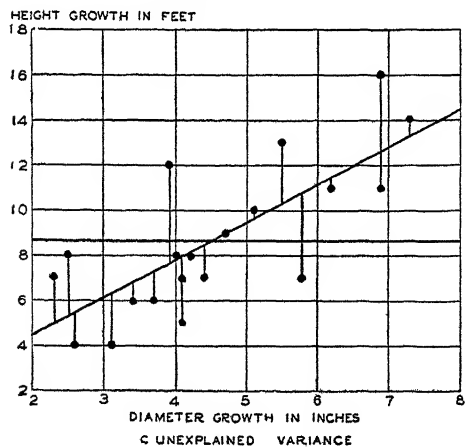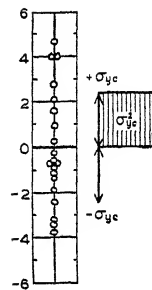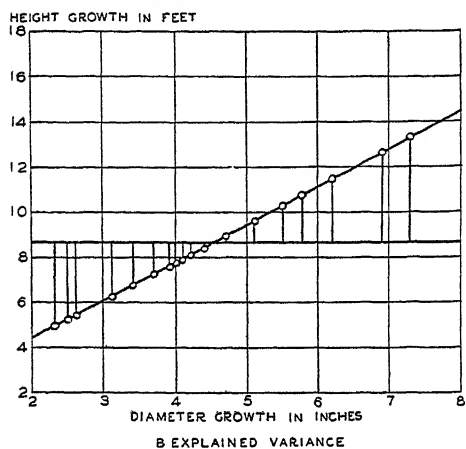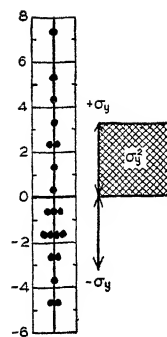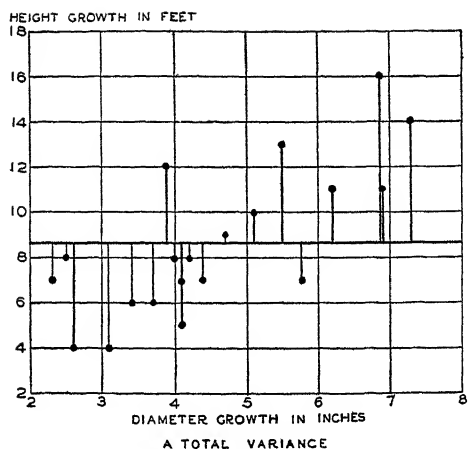
Chart 219. Derivation of Variance of Height Growth of 20 Forest Trees as Explained by Their Diameter Growth. (Data of Table 158.)

of the upper square—that is, total variance is the sum of the explained variance and unexplained variance, or

$$\sigma_y^2 = \sigma_{y_c}^2 + \sigma_{y_s}^2.$$

On the other hand, the standard deviation of the original data is smaller than the sum of the standard deviation of the computed values and the standard error of estimate. The relationship is clearly seen if we think of the standard deviation as being the hypotenuse of a right triangle and

EXPLAINED
VARIANCE

$\sigma_{y_c}^2 = 5.99$

UNEXPLAINED
VARIANCE

$\sigma_{y_s}^2 = 4.44$

$\sigma_{y_c} = 2.45$

$\sigma_{y_s} = 2.11$

$\sigma_y = 3.23$

TOTAL
VARIANCE

$\sigma_y^2 = 10.43$
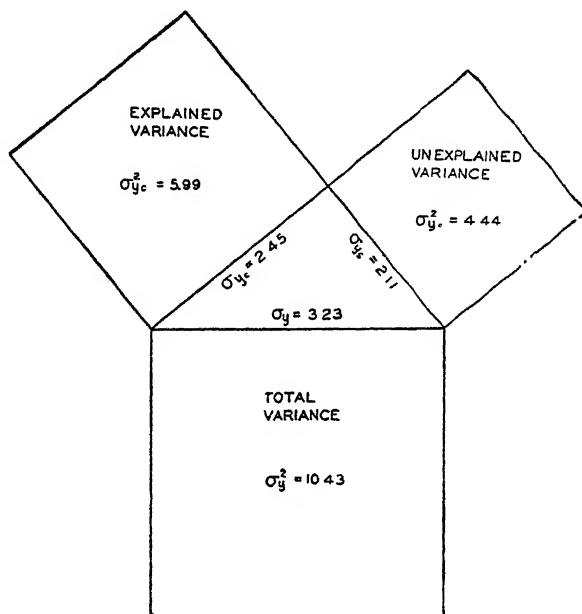
Chart 220. Diagrammatic Representation of Relationship between Standard Deviations and Variances. (Data of Table 158.)

the other two standard deviations as being the other two sides, as in Chart 220.[6]

The *coefficient of determination*,[7] r^2, is the proportion of total variance which has been explained (.574 in this case). The *coefficient of correlation,*

---

[6] See "Geometric Presentation of Correlation," by John W. Morse, *Journal of the American Statistical Association*, Vol. 32, June 1937, pp. 364–365. For proof that

$$\sigma_y^2 = \sigma_{y_c}^2 + \sigma_{y_s}^2$$

see Appendix B, section XXII–1, equation 9.

[7] The proportion of variance which has not been explained is sometimes called the coefficient of non-determination, k^2. From the table on p. 661, it can be seen that $r^2 + k^2 = 1$. Just as the square root of the coefficient of determination is the coefficient of correlation, so the square root of the coefficient of non-determination is known

r, is the square root of the coefficient of determination.[8] Thus the coefficient of correlation ($+ 758$) may be thought of as the square root of the proportion of variance that has been explained [9] r will, of course,

as the *coefficient of alienation*   But, although $r^2 + k^2 = 1$, $r + k > 1$ See Mordecai Ezekiel, *Methods of Correlation Analysis*, pp 376–377, John Wiley and Sons, New York, 1930

[8] Therefore r is also the ratio of the standard deviation of the computed values to the standard deviation of the original data; that is, $r = \dfrac{\sigma_{y_C}}{\sigma_y}$

A further simplification of the formula for r (of which use will be made in chapters to follow) is to regard r as the square root of the proportion of *variation* (sum of squared deviations) that has been explained   The development is as follows:

$$\sigma_y^2 = \sigma_{y_C}^2 + \sigma_{y_S}^2,$$

$$\frac{\Sigma(Y - \overline{Y})^2}{N} = \frac{\Sigma\, Y_C - \overline{Y})^2}{N} + \frac{\Sigma(Y - Y_C)^2}{N}.$$

$$\Sigma(Y - \overline{Y})^2 = \Sigma(Y_C - \overline{Y})^2 + \Sigma(Y - Y_C)^2.$$

Therefore r may be obtained from

$$r = \sqrt{\frac{\Sigma\, Y_C - \overline{Y})^2}{\Sigma(Y - \overline{Y})^2}}, \text{ or } \sqrt{1 - \frac{\Sigma(Y - Y_C)^2}{\Sigma(Y - \overline{Y})^2}}.$$

Since $Y - \overline{Y} = y$; $Y - Y_C = y_S$; $Y_C - \overline{Y} = y_C$, we may write

$$\Sigma y^2 = \Sigma y_C^2 + \Sigma y_S^2, \text{ and}$$

$$r = \sqrt{\frac{\Sigma y_C^2}{\Sigma y^2}}, \text{ or } \sqrt{1 - \frac{\Sigma y_S^2}{\Sigma y^2}}.$$

Occasionally a method of curve fitting is employed which results in lack of equality between Σy^2 and $\Sigma y_C^2 + \Sigma y_S^2$ For example, such a situation may obtain when the estimating equation is fitted by inspection.   In such cases it is customary to use the formula

$$r = \sqrt{1 - \frac{\sigma_{y_S}^2}{\sigma_y^2}}, \text{ or } \sqrt{1 - \frac{\Sigma y_S^2}{\Sigma y^2}}.$$

The reason for using this expression rather than the one involving Σy_C^2 is that the criterion of fit is the least squares criterion, and so the closeness of the correlation as well as the goodness of the fit depends on reducing the squares of the residuals from the fitted line.

[9] If the two variables X and Y are thought of as being composed of elements equally likely to be present in any item (some of which are common to X and Y, but some of which occur in the one and not the other), then the coefficient of determination of the entire population is the product of the two proportions of common elements, and the coefficient of correlation is their geometric mean.   Let us take 5 disks (elements) marked on one side as follows (the other side being blank):
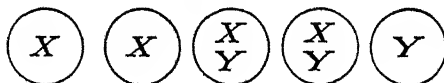
If we should throw all 5 disks in the air, when they fall any number of X's from 0 to 4 might appear, and also from 0 to 3 Y's.   Whenever an X appears, the chances that a

always be larger than r^2, unless $r^2 = 1$, when $r = r^2$.

$$r^2 = \frac{\sigma^2_{y_c}}{\sigma^2_y} = \frac{5.99}{10\ 43} = .574,$$

and

$$r = +.758.$$

Also

$$r^2 = \frac{\sigma^2_y - \sigma^2_{y_s}}{\sigma^2_y} = 1 - \frac{\sigma^2_{y_s}}{\sigma^2_y} = 1 - \frac{4.44}{10\ 43} = 1 - .426 = .574,$$

and

$$r = +.758.$$

Less laborious methods of computing r, which are not quite so straight forward in their meaning, are explained on pages 671–672.

The sign of r is always the same as the sign of b in the equation of relationship. Unless the value of the coefficient is very low, the sign of r can be determined by inspection of the scatter diagram.

The procedure just outlined has involved the obtaining of an equation with which to estimate values of Y from known values of X, and the explaining of variations in Y values by associating them with variations in X values. Mathematically it would be possible to obtain a line, the sum of the squares of the *horizontal* deviations from which is at a minimum, and from that line to explain variations in X values or to estimate X values from given Y values. The equation[10] $X_C = a' + b'Y$ would not describe the same line as $Y_C = a + bX$, nor would σ_{x_s} be the same as σ_{y_s}. However, r would be the same regardless of which route was used to obtain it. This is merely to say that mathematically, so far as r is concerned, either variable may be labeled X and the other Y. The X variable (plotted along the horizontal axis) is customarily the *causal* factor, if causation can logically be inferred; the Y variable is the *resultant* factor. If causation cannot be inferred, the factor which is the basis of our estimates is considered the X, or independent, variable. If, however, we wish to estimate values of the causal series from values of the resulting series, the causal

---

Y will also appear on the same disk are 2 out of 4; likewise, whenever a Y appears, the chances are 2 out of 3 that an X will appear on the same disk. If we should throw these disks in the air a number of times, counting the X's and Y's each time, there would be correlation between the number of X's that appear from throw to throw and the number of Y's The most likely value of r^2 is $\frac{2}{4} \times \frac{2}{3} = +.333$, while the most likely value of r is $\sqrt{\frac{2}{4} \times \frac{2}{3}} = + 58$ The larger the number of throws, the greater will be the tendency for r to approach this value. For a demonstration of this theory, see Croxton and Cowden, *Practical Business Statistics*, pp 416–419, Prentice Hall, Inc., New York, 1934.

[10] The normal equations required would be:

  I. $\Sigma X = Na' + b'\Sigma Y.$
  II. $\Sigma XY = a'\Sigma Y + b'\Sigma Y^2.$

factor may still be plotted as the X variable, but the estimating equation is of the type

$$X_C = a' + b'Y.$$

The product-moment formula. The coefficient of correlation may be approached from a number of different points of view. The approach which has been followed is especially enlightening, since essentially the same technique can be applied to curvilinear and multiple correlation. But the following explanation is also simple and, for certain purposes, extremely useful.

In the estimating equation b tells us the normal amount by which the dependent variable changes with a change of one unit in the independent variable. It is the slope or $\dfrac{y}{x}$ ratio of any point on the estimating equation, when y and x are defined as deviations from the mean of the series, so that the estimating equation becomes $y_C = bx$, and b is obtained by finding[11] the value of $\dfrac{\Sigma xy}{\Sigma x^2}$. Although this constant b is essential for purposes of estimation, still it cannot tell us the *degree* of relationship between the variables, since they are not directly comparable with each other. The X series and the Y series do not have the same dispersion, and may even be in different physical units. However, comparability between the terms of the ratio $\dfrac{y}{x}$ can be obtained by dividing the numerator by[12] σ_y and the denominator by σ_x or by dividing the entire expression by $\dfrac{\sigma_y}{\sigma_x}$. Thus, b is transformed into r as follows:

$$\frac{\Sigma xy}{\Sigma x^2} \div \frac{\sigma_y}{\sigma_x} = \frac{\Sigma xy}{\Sigma x^2}\frac{\sigma_x}{\sigma_y} = \frac{\Sigma xy}{N\sigma_x^2}\frac{\sigma_x}{\sigma_y} = \frac{\Sigma xy}{N\sigma_x\sigma_y}.$$

In this form the ratio is known as the *product-moment* form of the coefficient of correlation.[13] Thus it may be seen that r is merely the slope of

---

[11] See Appendix B, section XXII-2.

[12] Although we are referring to the standard deviation of the Y values, we use the symbol σ_y instead of σ_Y for convenience. It should be clear that $\sigma_Y = \sigma_y$, because $\dfrac{\Sigma y}{N} = 0$, $Y - \overline{Y} = y - 0$, and $\Sigma(Y - \overline{Y})^2 = \Sigma y^2$.

[13] Another way of getting the same result is to think of r as a special case of b; namely, when the original data have been made comparable by expressing them in units of their own standard deviations. Thus

$$\frac{\Sigma xy}{\Sigma x^2} \text{ becomes } \frac{\Sigma\left(\dfrac{x}{\sigma_x}\right)\left(\dfrac{y}{\sigma_y}\right)}{\Sigma\left(\dfrac{x}{\sigma_x}\right)^2} = \frac{\Sigma xy}{\sigma_x\sigma_y}\frac{\sigma_x^2}{\Sigma x^2} = \frac{\Sigma xy}{\sigma_x\sigma_1}\frac{\sigma_x^2}{N\sigma_x^2} = \frac{\Sigma xy}{N\sigma_x\sigma_y}.$$

the estimating equation when both numerator and denominator are in standard deviation units.

Now since

$$r = b \div \frac{\sigma_y}{\sigma_x},$$

$$b = r \frac{\sigma_y}{\sigma_x},$$

and

$$y_C = r \frac{\sigma_y}{\sigma_x} x.$$

Analogously

$$x_C = r \frac{\sigma_x}{\sigma_y} y.$$

Use of the estimating equation in this form, $y_C = r \dfrac{\sigma_y}{\sigma_x} x$, will be made later in this chapter.

Practical Methods of Computation

The previous illustration involved a limited number of paired items in order to illustrate the theory of correlation as concisely as possible. In most practical problems, however, we have a large number of pairs of items. In practice, therefore, it is advisable to modify the foregoing methods slightly in order to save time, and we shall illustrate the shorter procedures by means of a sample involving 64 pairs of items. The reader can readily see that the procedure described previously would be very laborious when applied to such a problem.

As the initial step in a correlation problem, a scatter diagram should always be drawn. If only an approximate idea of the degree of relationship is required, inspection of the scatter plot yields satisfactory results. After a little experience in correlating, the statistician is able to make surprisingly close estimates of r by inspection. The scatter diagram may frequently be used for exploratory purposes and may occasionally yield sufficient information to eliminate the need of determining the coefficient of correlation.

The formula is often stated also as $r = \dfrac{1}{N} \Sigma \left(\dfrac{x}{\sigma_x} \cdot \dfrac{y}{\sigma_y} \right)$. The reason for the adjective "product-moment" becomes clear when it is realized that the word "moment" refers to the average of some power of the deviations from a mean. Thus r is the first moment of the product of the variables when each has been previously stated in terms of its own standard deviation. For proof that $\dfrac{\Sigma xy}{N \sigma_x \sigma_y} = \sqrt{\dfrac{\Sigma y_C^2}{\Sigma y^2}}$, see Appendix B, section XXII–3.

The data used in the following illustration of procedure are dividends per share and lowest price per share, during 1935, of common stocks of 64 American industrial corporations. These companies were selected at random from *Moody's Industrials, 1936*. Chart 221 is the scatter diagram, while the data and computations required for values used in the formulae are shown in Table 159.
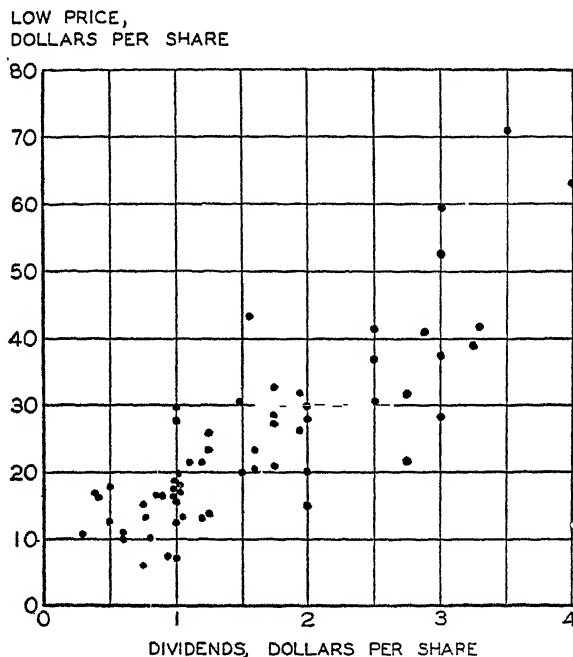
Chart 221.   Dividends per Share and Low Price per Share of 64 Medium Grade Common Stocks of American Industrial Corporations, 1935.   (Data of Table 159.)

Substituting in the two normal equations

$$\Sigma Y = Na + b\Sigma X,$$
$$\Sigma XY = a\Sigma X + b\Sigma X^2,$$

we have:

$$1{,}600.61 = 64a + 102.94b,$$
$$3{,}198.1790 = 102.94a + 216.9558b,$$

and the equations solved simultaneously yield the estimating equation

$$Y_C = 5.48606 + 12.1382X.$$

Values for a and b have been carried to six digits, since subsequent calculations will be based on them.

TABLE 159

COMPUTATION OF VALUES USED IN DETERMINING MEASURES OF RELATIONSHIP BE-
TWEEN DIVIDENDS PER SHARE AND LOW PRICE IN 1935 OF MEDIUM GRADE
STOCKS OF 64 AMERICAN CORPORATIONS

Company*	Dividends per share in 1935 X	Low price during 1935 Y	XY	X^2	Y^2
Alaska Juneau Mining Co	$1.20	$13 25	15 9000	1 4400	175 5625
American Agric Chem Co (Det)	2 50	41.50	103 7500	6 2500	1,722 2500
American Machinery & Foundry Co	1 00	18 50	18 5000	1 0000	342 2500
Anchor Cap Corp .	.60	10 88	6 5280	3600	118 3744
Armstrong Cork Co	88	16 50	14 5200	7744	272 2500
Associated Oil Company	1 00	29 75	29.7500	1.0000	885 0625
Bloomingdale Brothers, Inc	40	16 62	6.6480	1600	276 2244
Borg-Warner Corp	1 75	28 25	49 4375	3 0625	798 0625
Brown Shoe Co	3 00	53 00	159 0000	9 0000	2,809 0000
Burroughs Adding Machine Co	. 1 05	13 25	13 9125	1 1025	175 5625
California Packing Corp ...	1 50	30 50	45 7500	2 2500	930 2500
Cannon Mills Co .	2 00	30 00	60 0000	4 0000	900 0000
Chicago Mail Order Co .	2.00	15.12	30 2400	4 0000	228 6144
Cleveland Graphite Bronze Co . . .	1 25	27 62	34 5250	1 5625	762 8644
Colgate-Palmolive Peat Co	.75	15 12	11 3400	.5625	228 6144
Commercial Solvents Corp.	85	16.50'	14 0250	.7225	272 2500
Cudahy Packing Co . .	2 50	37 00	92 5000	6 2500	1,369 0000
Diamond Match Co. ..	1 95	26 50	51 6750	3 8025	702 2500
Duplan Silk Corp .	1 00	12 75	12 7500	1 0000	162 5625
Electric Storage Battery Co	3.25	39 00	126.7500	10 5625	1,521.0000
Eureka Vacuum Cleaning Co , Inc	80	10 50	8 4000	6400	110 2500
Freeport Texas Co .	1 00	17.25	17.2500	1 0000	297.5625
General American Trans- portation Co.	1 75	32 62	57.0850	3 0625	1,064 0644
General Mills, Inc...	3.00	59 88	179.6400	9 0000	3,585.6144
Glidden Company .. .	1.60	23.38	37.4085	2.5600	546.6244
Granite City Steel Co. ..	1.00	18 12	18.1200	1.0000	328.3344
W. T. Grant Co. ..	1.25	26.00	32.5000	1 5625	676 0000
Harbison-Walker Refracto- ries Co..	1.00	16.00	16.0000	1.0000	256.0000
Hercules Powder Co ..	3.50	71.00	248.5000	12 2500	5,041.0000
Industrial Rayon Corpora- tion....	1.26	23.50	29 6100	1 5876	552.2500
International Printing Ink Corp	1 10	21.50	23.6500	1.2100	462.2500
Kaufman Department Stores, Inc.	1 00	7.50	7.5000	1.0000	56 2500
S. S. Kresge Co.	1 00	19.75	19 7500	1 0000	390 0625
Lambert Co.. . .	2 75	21.38	58 7950	7.5625	457.1044
Libby-Owens-Ford Glass Co	1 20	21.50	25 8000	1.4400	462 2500

TABLE 159 (Continued)

COMPUTATION OF VALUES USED IN DETERMINING MEASURES OF RELATIONSHIP BE-
TWEEN DIVIDENDS PER SHARE AND LOW PRICE IN 1935 OF MEDIUM GRADE
STOCKS OF 64 AMERICAN CORPORATIONS

Company*	Dividends per share in 1935 X	Low price during 1935 Y	XY	X^2	Y^2
Manhattan Shirt Co	$ 60	$10 00	6 0000	.3600	100 0000
Marlin-Rockwell Corp .	1 50	20 00	30 0000	2 2500	400 0000
McCall Corporation . .	2 00	28 00	56 0000	4 0000	784 0000
Melville Shoe Corporation.	2 88	41 00	118 0800	8 2944	1,681 0000
MacAndrews and Forbes Co	3 00	37 88	113 6400	9 0000	1,434 8944
John Morrell and Co , Inc	3.30	41.88	138 2040	10 8900	1,753 9344
Natomas Co.	.95	7 50	7 1250	9025	56 2500
Newberry (J J) Co .	1.45	43 50	63 0750	2 1025	1,892 2500
Peoples Drug Stores, Inc .	2.75	30 00	82 5000	7 5625	900 0000
Phelps Dodge Corp . . .	50	12 75	6 3750	2500	162 5625
Pillsbury Flour Mills Co. .	1.60	31.00	49 6000	2 5600	961 0000
Phillips Petroleum Co. .	1.25	13 75	17 1875	1 5625	189 0625
Pullman Incorporated	2.62	29 50	77 2900	6 8644	870 2500
Raybestos Manhattan, Inc.	1 00	16.50	16 5000	1 0000	272 2500
Reynolds Metals Co.. . .	1 00	17.50	17 5000	1 0000	306 2500
Safeway Stores, Inc. . . .	2 75	31 62	86 9550	7 5625	999 8244
Socony-Vacuum Oil Co. . .	30	10 62	3 1860	0900	112 7844
South Porto Rico Sugar Co.	2 00	20 00	40 0000	4 0000	400 0000
Spencer Kellogg & Sons Co.	1 60	31 00	49.6000	2 5600	961 0000
Standard Oil Co. of California	1 00	27 75	27.7500	1 0000	770 0625
Telautograph Corp . . .	75	6.25	4.6875	5625	39 0625
Thatcher Manufacturing Co	.75	13 12	9 8400	5625	172 1344
Timken Roller Bearing Co .	3.00	28 38	85 1400	9 0000	805 4244
United Biscuit Co. of America	1.60	20 25	32 4000	2 5600	410 0625
Vulcan Detinning Co . . .	4 00	63 50	254 0000	16 0000	4,032 2500
Warren Foundry & Pipe Corp.	1.75	20 62	36 0850	3 0625	425 1844
Wesson Oil & Snowdrift Co., Inc	2 50	30 50	76 2500	6 2500	930 2500
Westinghouse Air Brake Co.	50	18 00	9 0000	.2500	324 0000
Westvaco Chlorine Products Corp.	40	16 75	6 7000	.1600	280 5625
Total	102 94	1,600 61	3,198.1790	216 9558	51,363 9273

Source Moody's *Industrials, 1936*
* Corporations were selected at random from a list appearing in the *Annalist* The sample includes
only corporations whose shares were listed on the New York Stock Exchange and traded in during 1935
It includes only stocks with Fitch ratings of B and BB, and which paid dividends in 1934 and 1935

Instead of computing σ_{y_s} from the formula

$$\sigma_{y_s}^2 = \frac{\Sigma y_s^2}{N} = \frac{\Sigma(Y - Y_c)^2}{N},$$

a less direct method is much easier. First we compute ΣY_c^2 by the expression

$$\Sigma Y_c^2 = a\Sigma Y + b\Sigma XY.$$

Then we obtain[14]

$$\begin{aligned}\Sigma y_s^2 &= \Sigma Y^2 - \Sigma Y_c^2 \\ &= \Sigma Y^2 - (a\Sigma Y + b\Sigma XY).\end{aligned}$$

The indirect but very easy formula for obtaining σ_{y_s}, then, is

$$\sigma_{y_s}^2 = \frac{\Sigma Y^2 - (a\Sigma Y + b\Sigma XY)}{N}.$$

Substituting in this formula, we find:

$$\sigma_{y_s}^2 = \frac{51{,}363.9273 - [(5.48606)(1{,}600.61) + (12.1382)(3{,}198.1790)]}{64}$$

$$= \frac{51{,}363.93 - 47{,}601.18}{64} = 58.79,$$

and $\sigma_{y_s} = \$7.667$.

The coefficient of correlation may be obtained as follows:[15]

$$\begin{aligned}r^2 = \frac{\sigma_{y_c}^2}{\sigma_y^2} &= \frac{\Sigma y_c^2 \div N}{\Sigma y^2 \div N} = \frac{\Sigma y_c^2}{\Sigma y^2} \\ &= \frac{(a\Sigma Y + b\Sigma XY) - \bar{Y}\Sigma Y}{\Sigma Y^2 - \bar{Y}\Sigma Y}.\end{aligned}$$

This is a very easy expression to use, since most of the values have already been computed. Thus

$$r^2 = \frac{47{,}601.18 - (25.0095)(1{,}600.61)}{51{,}363.93 - (25.0095)(1{,}600.61)} = \frac{7{,}570.543}{11{,}333.422} = .6680,$$

and

$$r = +.8173.$$

It is also easy to obtain r by the formula:

$$r^2 = 1 - \frac{\sigma_{y_s}^2}{\sigma_y^2}.$$

---

[14] Proof of the formulae for ΣY_c^2 and Σy_s^2 is given in Appendix B, section XXII-1, equations 3, 6, and 7.

[15] For proof that $\Sigma y_c^2 = (a\Sigma Y + b\Sigma XY) - \bar{Y}\Sigma Y$, see Appendix B, section XXII-1, equations 4 and 5 For proof that $\Sigma y^2 = \Sigma Y^2 - \bar{Y}\Sigma Y$, see Appendix B, section XIII-2, part B-1

The value of $\sigma_{y_s}^2$ has already been computed, and ΣY^2 and ΣY are available, from which to compute σ_y^2. The sign of r is not given by these expressions, but r takes the sign of b in the estimating equation, or, as noted earlier, the sign may usually be discovered from an examination of the scatter diagram.

If it is desired to compute r first, and to regard σ_{y_s} and the estimating equation as a by-product, it may be done quite readily. It will be recalled that one expression for the coefficient of correlation mentioned on page 666 is

$$r = \frac{\Sigma xy}{N\sigma_x\sigma_y}.$$

In this formula the variables are taken as deviations from their respective means   If they are taken in their original form ($x = X - \overline{X}$ and $y = Y - \overline{Y}$), then[16]

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}}.$$

Thus

$$r = \frac{(64)(3,198.1790) - (102\ 94)(1600.61)}{\sqrt{[(64)(216.9558) - (102.94)^2][(64)(51,363.927) - (1600.61)^2]}}$$
$$= +.8173.$$

Having obtained r, we may compute our estimating equation by means of the two normal equations used before, or by use of the equation developed on page 667:

$$y_c = r\frac{\sigma_y}{\sigma_x}x.$$

Since the data are not in deviation form, the equation must be written

$$(Y_C - \overline{Y}) = r\frac{\sigma_y}{\sigma_x}(X - \overline{X}).$$

The equation for estimating price from dividends therefore becomes

$$Y_C - 25.0095 = .8175\frac{13.3079}{.896047}(X - 1.6084),$$

and

$$Y_C = 5.4861 + 12.138X,$$

which is the same as previously obtained.

The calculation of the means and standard deviations, which this pro-

---

[16] For derivation of this formula, see Appendix B, section XXII–4   The value of r^2 may be obtained by squaring r. However, a somewhat more accurate figure may be obtained by first squaring the numerator and denominator of the expression for r

cedure requires, is not shown, since it is a matter with which the student is already acquainted.

Correlation of Grouped Data

Commodities differ greatly in the frequency with which they change their price. Some, such as agricultural implements, are very rigid and change only after mature deliberation on the part of sellers, while others, like commodities on an organized exchange, are very flexible, changing frequently from minute to minute. During the depression which began in 1929, it was noticed that some of those articles, the prices of which seldom changed, failed to drop so far as those with greater price flexibility.

Was this a general tendency? Is there a close correlation between price flexibility and price change? With over 700 price series available, collected by the United States Bureau of Labor Statistics, this question can be answered rather definitely. But since the pairs of items to be correlated are large in number, it is easier to group them before undertaking calculations. First the data are tallied as in Table 160, which resembles a scatter diagram except that each point, instead of being plotted exactly, is merely entered in the appropriate cell. Thus a commodity which changed in price during 93 of the 94 months under consideration (1926–1933), and also declined during the period 1929–1932 to 18 per cent of its 1929 price, would be tallied in the extreme lower right-hand corner.[17]

Table 161 is a correlation table. The figures in the center of each cell are taken from Table 160. The f_Y values are obtained by adding the numbers horizontally; the f_X values, by adding vertically. These two sets of figures will be recognized as frequency distributions of the dependent and independent variables respectively. The total frequencies, or commodities N, for each distribution are, of course, the same: 736. The three other columns and rows in the table are identical with those to which we are accustomed for computing the mean and standard deviation from a frequency distribution, except that here we have two frequency distributions, one of the X values (running horizontally) and another of the Y values (running vertically). For ease in computation, deviations are measured in terms of class intervals from assumed means, that of X being chosen as 44.5 price changes and that of Y as 65 per cent of 1929.

Since XY values are required for r, these also are computed for each cell and totaled. This is done by multiplying the X deviation by the Y

---

[17] The data are taken from a scatter diagram which appeared on p. 3 of Gardiner C. Means, "Industrial Prices and Their Relative Inflexibility," *Senate Document No. 13, 74th Congress 1st Session, 1935.* The data on flexibility run from January 1926 through December 1933, except that no observation was taken for the period December 1929–January 1930.

deviation (shown in the upper part of each cell), and finally multiplying this product by the appropriate frequency. The results are shown in boldface type in the lower part of each cell. It will be noticed that the first and third quadrants are positive, while those in the second and fourth are, of course, negative. The algebraic total of these products is shown in the lower right-hand corner of the table. There is no subscript for f in

TABLE 160

TABULATION OF PRICE FLEXIBILITY AND MAGNITUDE OF PRICE CHANGE OF
736 COMMODITIES

MAGNITUDE OF PRICE CHANGE (Y)
PER CENT OF 1929

The following table gives the boldface (circled) frequency shown in the lower part of each cell.

Y range	0–9	10–19	20–29	30–39	40–49	50–59	60–69	70–79	80–89	90–99
160.0–169.9	(1)									
150.0–159.9	(1)									
140.0–149.9										
130.0–139.9										
120.0–129.9										
110.0–119.9	(6)	(3)		(2)	(2)					
100.0–109.9	(43)	(7)		(2)	(0)				(0)	(0)
90.0–99.9	(74)	(4)	(2)	(0)		(0)				
80.0–89.9	(36)	(35)	(10)	(13)	(2)	(2)	(0)			
70.0–79.9	(23)	(30)	(17)	(12)	(8)	(2)	(3)	(0)	(2)	(4)
60.0–69.9	(10)	(29)	(21)	(14)	(6)	(9)	(7)	(3)	(9)	(9)
50.0–59.9	(2)	(9)	(12)	(8)	(5)	(9)	(9)	(9)	(11)	(28)
40.0–49.9		(2)	(0)	(3)	(0)	(4)	(7)	(6)	(21)	(44)
30.0–39.9		(0)			(0)	(2)	(4)	(3)	(9)	(25)
20.0–29.9								(0)	(2)	(3)
10.0–19.9										(2)

PRICE FLEXIBILITY (X)
NUMBER OF PRICE CHANGES, 1926–1933

the expression $\Sigma f d'_X d'_Y$ since each cell frequency is common to an X class and to a Y class.

It would be possible now to set up two normal equations and obtain the estimating equation directly, as was done for the ungrouped data in other parts of this chapter.[18] Such a procedure would be necessary for certain

[18] The normal equations are:

$$\text{I}\qquad \Sigma f_Y d'_Y = Na + b\Sigma f_X d'_X;$$
$$\text{II.}\quad \Sigma f d'_X d'_Y = a\Sigma f_X d'_X + b\Sigma f_X (d'_X)^2.$$

Making the necessary substitutions, we have·

$$\text{I.}\qquad 462 = 736a - 422b;$$
$$\text{II.}\qquad -4317 = 422a + 8796b$$

Solved simultaneously, these yield the estimating equation

$$d'_{Y_C} = 3561 - .4737\, d'_X.$$

The following table explains the computation of Y_C values from this equation·

X	d'_X	d'_{Y_C} $(a + b\, d'_X)$	d_{Y_C} $(i_Y \cdot d'_{Y_C})$	Y_C $(\overline{Y}_d + d_{Y_C})$
4 5	−4	2 2509	22 509	87.51
24 5	−2	1 3035	13.035	78.04
44 5	0	.3561	3 561	68.56
64 5	2	− .5913	− 5.913	59 09
84 5	4	−1.5387	−15.387	49 62

Computation of Y_C values other than mid-values of classes, however, is facilitated by stating the equation in terms of the original data. When so stated, the equation is $Y_C = 89.64 - 4737X$. The procedure for making this transformation is explained in Appendix B, section XXII–5

We can now find σ_{y_s} and r by use of formulae paralleling those used in ungrouped data.

$$\sigma^2_{y_s} = \frac{\Sigma f_Y (d'_Y)^2 - (a\Sigma f_Y d'_Y + b\Sigma f d'_X d'_Y)}{N}$$

$$= \frac{3822 - [.3561(462) - .4737(4317)]}{736}$$

$$= \frac{3822 - 2209\ 5070}{736} = 2\ 1909.$$

$$\sigma_{y_s} = \sqrt{2.1909} = 1\ 4802 \text{ intervals} = 14.802 \text{ per cent.}$$

$$r^2 = \frac{N(a\Sigma f_Y d'_Y + b\Sigma f d'_X d'_Y) - (\Sigma f_Y d'_Y)^2}{N\Sigma f_Y (d'_Y)^2 - (\Sigma f_Y d'_Y)^2}$$

$$= \frac{736(2209.5070) - (462)^2}{736\ (3822) - (462)^2} = .5435.$$

$$r = \sqrt{5435} = -.7372.$$

TABLE 161

Correlation Table of Price Flexibility and Magnitude of Price Change of 736 Commodities

PRICE FLEXIBILITY (X)
Number of price changes, 1926-1933.

MAGNITUDE OF PRICE CHANGE (Y) — Per Cent of 1929

Class limits	Mid-value	0–9 (4 5)	10–19 (14 5)	20–29 (24 5)	30–39 (34 5)	40–49 (44 5)	50–59 (54 5)	60–69 (64 5)	70–79 (74 5)	80–89 (84 5)	90–99 (94 5)	f_Y	d'_Y	$f_Y d'_Y$	$f_Y(d'_Y)^2$
160 0–169 9	165	−40 / 1 / −40										1	10	10	100
150 0–159 9	155	−36 / 1 / −36										1	9	9	81
140 0–149 9	145											8	0	0	0
130 0–139 9	135											7	0	0	0
120 0–129 9	125											6	0	0	0
110 0–119 9	115	−20 / 6 / −120	−15 / 3 / −45		−5 / 2 / −10	0 / 2 / 0						13	5	65	325
100 0–109 9	105	−16 / 43 / −688	−12 / 7 / −84		−4 / 2 / −8	0 / 1 / 0				+16 / 1 / 16	+20 / 1 / 20	55	4	220	880
90 0–99 9	95	−12 / 74 / −888	−9 / 16 / −144	−6 / 2 / −12	−3 / 10 / −30		+3 / 1 / 3					103	3	309	927
80 0–89.9	85	−8 / 36 / −288	−6 / 35 / −210	−4 / 10 / −40	−2 / 13 / −26	0 / 2 / 0	+2 / 2 / 4	+4 / 1 / 4				99	2	198	396
70 0–79.9	75	−4 / 23 / −92	−3 / 30 / −90	−2 / 17 / −34	−1 / 12 / −12	0 / 8 / 0	+1 / 2 / 2	+2 / 3 / 6	+3 / 1 / 3	+4 / 2 / 8	+5 / 4 / 20	102	1	102	102
60 0–69.9	65	0 / 10 / 0	0 / 29 / 0	0 / 21 / 0	0 / 14 / 0	0 / 6 / 0	0 / 9 / 0	0 / 7 / 0	0 / 3 / 0	0 / 8 / 0	0 / 9 / 0	116	0	0	0
50 0–59 9	55	+4 / 4 / 16	+3 / 9 / 27	+2 / 12 / 24	1 / 8 / 8	0 / 5 / 0	−1 / 9 / −9	−2 / 9 / −18	−3 / 9 / −27	−4 / 11 / −44	−5 / 28 / −140	104	−1	−104	104
40 0–49 9	45		+6 / 2 / 12	4 / 1 / 4	2 / 3 / 6	0 / 1 / 0	−2 / 4 / −8	−4 / 7 / −28	−6 / 6 / −36	−8 / 21 / −168	−10 / 44 / −440	89	−2	−178	356
30 0 39 9	35		9 / 1 / 9			0 / 1 / 0	−3 / 2 / −6	−6 / 4 / −24	−9 / 3 / −27	−12 / 9 / −108	−15 / 25 / −375	45	−3	−135	405
20 0 29 9	25							−12 / 1 / −12		−16 / 2 / −32	−20 / 3 / −60	6	−4	−24	96
10 0 19 9	15										−25 / 2 / −50	2	−5	−10	50
f_X		198	132	63	64	26	29	31	23	54	116	N = 736		$\Sigma f_Y d'_Y$ = 462	$\Sigma f_Y(d'_Y)^2$ = 3,822
d'_X		−4	−3	−2	−1	0	1	2	3	4	5				
$f_X d'_X$		−792	−396	−126	−64	0	29	62	69	216	580	$\Sigma f_X d'_X$ = −422		$\Sigma f d'_X d'_Y$ = −4,317	
$f_X(d'_X)^2$		3168	1188	252	64	0	29	124	207	864	2900	$\Sigma f_X(d'_X)^2$ = 8,796			

Source: Table 160.

non-linear estimating equations.   However, it is somewhat simpler to compute first the correlation coefficient, and then the estimating equation and standard error of estimate.

To obtain r directly from ungrouped data the following formula has been recommended:

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}}.$$

For grouped data, X is replaced by d'_X and Y by d'_Y, the symbol f is introduced, and the expression becomes

$$r = \frac{N\Sigma fd'_X d'_Y - (\Sigma f_X d'_X)(\Sigma f_Y d'_Y)}{\sqrt{[N\Sigma f_X(d'_X)^2 - (\Sigma f_X d'_X)^2][N\Sigma f_Y(d'_Y)^2 - (\Sigma f_Y d'_Y)^2]}}.$$

Substituting in this formula, we have

$$r = \frac{(736)(-4317) - (-422)(462)}{\sqrt{[(736)(8796) - (-422)^2][(736)(3822) - (462)^2]}}$$

$$= -.7372.$$

The following measures are readily computed by familiar methods:

$$\bar{X} = 38.766. \qquad \bar{Y} = 71.277.$$
$$\sigma_x = 34.090. \qquad \sigma_y = 21.906.$$

Now since

$$r^2 = 1 - \frac{\sigma^2_{y_s}}{\sigma^2_y},$$

$$\sigma^2_{y_s} = \sigma^2_y(1 - r^2),$$

$$\sigma_{y_s} = \sigma_y\sqrt{1 - r^2}.$$

Substituting:

$$\sigma_{y_s} = 21.906\sqrt{1 - (-.7372)^2}$$

$$= 14.802.$$

To obtain the estimating equation, we have the equation $y_s = r\dfrac{\sigma_y}{\sigma_x}x$. But since $y = Y - \bar{Y}$, and $x = X - \bar{X}$, we know that

$$Y_c - \bar{Y} = r\frac{\sigma_y}{\sigma_x}(X - \bar{X}).$$

Substituting in this equation, we have

$$Y_c - 71.277 = -.7372\frac{21.906}{34.090}(X - 38.766),$$

or

$$Y_c = 89.64 - .4737X.$$

These values are not exactly the same as would have been obtained had the computations been based upon ungrouped data. The difference, however, is ordinarily very slight and is due to the fact that the items are not distributed evenly within each cell, and hence the mid-values do not correspond with the actual means of each cell. The errors tend to offset each other, provided the X and Y distributions are approximately symmetrical. But this is merely a tendency and there is almost always a small discrepancy in the results. In general it may be said that, in order to keep the inaccuracy within negligible limits, it is well to have at least twelve groups in each direction.

Causation and the Correlation Coefficient

The coefficient of correlation must be thought of, not as something that proves causation, but only as a measure of co-variation. Any one of the following situations may, in fact, obtain:

1. *A variation in one variable may be caused (directly or indirectly) by a variation in the other.* The variable that is supposed to be the cause of variations in the other is usually taken as the independent variable and plotted along the X-axis. Thus, because dividends on stocks are thought to affect stock prices, rather than vice versa, the "dividends" series was made the independent variable, in an earlier illustration. It is a logical process which determines the statistician's belief that there is causal relationship between the two variables, and his belief as to which is cause and which is effect. It must be evident, then, that the coefficient of correlation in itself does not say that X causes Y, any more than it says that Y causes X.

2. *Co-variation of the two variables may be due to a common cause or causes affecting each variable in the same way, or in opposite ways.* If it should be found that there is correlation between automobile accidents per 1,000 persons and per capita federal income tax payments, it should not hastily be concluded that it takes an automobile accident to jar a person into paying his income tax; nor is it necessarily true that making large tax payments incapacitates a person for driving carefully. It is quite possible, however, that in states where the average income is high, the income taxes will be large, a large proportion of the people will own automobiles, and accidents will be numerous.

3. *The causal relationship between the two variables may be interacting.* Thus, a high price for a commodity stimulates its production, but increased production may increase or decrease the cost of a commodity, depending upon the period of time under observation and whether it is an increasing or decreasing cost industry.

4. *The correlation may be due to chance.* Even though there may be no relationship whatever between the variables in the universe from which the sample is drawn, it may be that enough of the paired variables that are selected may vary together, just by chance, to give a fair degree of correlation. Thus it might be found that in a given group of male students there was positive correlation between the size of their shoes and the number of cigarettes in their pockets. Yet it is hard to develop a theory as to why this should be so, and the chances are that another sample would yield quite different results. In a later section brief attention will be given to measurement of the reliability of r.

Estimate of Correlation in Population

All of the computations so far have measured the correlation in the particular sample. This tends to be higher than the correlation in the population from which the sample was drawn, and especially so when the sample is small. Estimating the correlation in the population may therefore be thought of as making allowance for the size of the sample. The best estimate of correlation in the population is from the formula[19]

$$\bar{r}^2 = 1 - \frac{\bar{\sigma}_{y_s}^2}{\bar{\sigma}_y^2} = \frac{r^2\,(N-1) - (m-1)}{N-m},$$

in which m is the number of constants in the estimating equation (always 2 in case of simple linear correlation). The formula therefore simplifies to

$$\bar{r}^2 = \frac{r^2(N-1) - 1}{N-2}.$$

Applying this correction to the forest tree illustration, we have

$$\bar{r}^2 = \frac{.574(19) - 1}{18} = .550$$

$$\bar{r} = \sqrt{.550} = +.742.$$

This compares with the sample r of $+.758$. A similar correction when applied to the common stock problem, a sample of 64 items, lowers the correlation only from $+.817$ to $+.814$.

If the value of r^2 is very low, $\bar{r}^2$ may be negative (and $\bar{r}$ imaginary) In such a case the correlation in the population should be considered to be zero.

---

[19] This formula is derived in Appendix B, section XXII-6.

Reliability of the Correlation Coefficient

General measure of reliability. The standard error of the correlation coefficient (σ_r), which is analogous to the standard error of the mean, is computed from the expression

$$\sigma_r = \frac{1 - r_P^2}{\sqrt{N - 1}},$$

where r_P is the correlation in the population.

This measure of the sampling error of r is subject to certain limitations. In the first place, the distribution of sample coefficients around the population r is approximately normal only in case the latter is zero. When the population r is positive, the sampling distribution is negatively skewed. Thus, if the true r is $+.80$, different sample coefficients can be only .20 higher, but some might conceivably be as low as -1.00 (a drop of -1.80). As r_P approaches zero, the distribution gradually approaches normality. No precise line of demarcation can be drawn beyond which it is unwise to use σ_r, but it has been suggested by Tippett that consideration of skewness becomes especially important as r_P approaches .80. A second limitation is that, even when r_P is zero or nearly so, the distribution of r is not normal for small samples, but approaches normality as the sample size is increased.

We are sometimes interested in testing whether or not there is *any* significant correlation; that is, whether the hypothesis is tenable that there is no correlation present in the population. If the hypothesis is discredited, the correlation is considered significant. To test this hypothesis, we must substitute zero for r_P in the formula for σ_r, which now becomes

$$\sigma_r = \frac{1}{\sqrt{N - 1}}.$$

This expression, as indicated in the preceding paragraph, should not be used when N is small.

It will be recalled that a random sample of 64 corporations showed a correlation coefficient between dividends and price of $+.8173$. Using the formula above,

$$\sigma_r = \frac{1}{\sqrt{63}} = .1260.$$

As is usual when testing for significance, we must divide the difference to be tested by the appropriate standard error. Thus we have

$$\frac{r}{\sigma_r} = \frac{.8173}{.1260} = 6.49.$$

Since statisticians usually have considerable confidence that a relationship

is not due to chance if r is at least three times σ_r, clearly our correlation is significant. If the test indicates lack of significance, the remedy is to increase the size of the sample.

The expression

$$\sigma_r = \frac{1 - r^2}{\sqrt{N - m}},$$

where m is the number of constants in the estimating equation, is sometimes used as a test of the reliability of r. For the 64 corporations

$$\sigma_r = \frac{1 - .6680}{\sqrt{62}} = .042.$$

The correlation coefficient is then written $r = +.817 \doteq .042$, which is interpreted to mean that, if the correlation in the population were $+.817$, 68.27 per cent of the coefficients computed from random samples of 64 pairs of items would be expected to vary between $+.772$ and $+.859$. This is a very crude and unsatisfactory procedure, however, since $+.817$ is not the population figure and, even if the value of r_P were $+.817$, the distribution of sample r's around such an r_P would not be normal. The distribution of sample r's would be approximately normal only if N were large and r_P were small. Using $-3\sigma_r$, it is sometimes asserted that it is unlikely that the value of r_P is below $+.817 - 3(.042) = +.691$. This, however, is an extremely rough application of fiducial probability. More satisfactory fiducial limits of r may be obtained by transforming r into Z (discussed later in this chapter), ascertaining the desired fiducial limits of Z, and converting to r.

The t test. In order to discover whether or not an observed correlation coefficient is significantly greater than zero, we may use a procedure which is applicable to both large and small samples. This method consists in computing the value t from the expression

$$t = r \div \frac{\sqrt{1 - r^2}}{\sqrt{N - m}} = \frac{r\sqrt{N - m}}{\sqrt{1 - r^2}} = \frac{r\sqrt{N - 2}}{\sqrt{1 - r^2}},$$

where m is the number of constants in the estimating equation. Following this we consult the t table of Appendix F (described in Chapter XII on unreliability), referring to the values of t and of n ($n = N - 2$), and discover how many times in 100 a sample drawn from a population with zero correlation would result in a correlation coefficient as high as that actually obtained. If this chance is very low, the correlation is assumed to be significant. For the illustration just discussed,

$$t = \frac{.8173\sqrt{62}}{\sqrt{1 - .6680}} = 11.2.$$

Since t gives a ratio of 11.2 and $n = 62$, it appears that a correlation coefficient of this size could hardly have been due to chance if based on a sample drawn from a population having zero correlation, and furthermore that, if we should correlate yield and price of all common stocks such as these, we should find positive correlation to obtain.

The value of $\frac{r}{\sigma_r}$, when $\sigma_r = \frac{1}{\sqrt{N-1}}$, was found to be 6.49 as compared with a value for t of 11.2. Now, the t distribution is almost normal when n exceeds 30. Since n is 62 in this example, it is apparent that the test involving σ_r errs on the side of stringency when testing for presence of correlation, since this test requires a larger value of r to give a ratio which would indicate the same probability as that found by the t test. The t test is the more appropriate procedure to use when testing for the presence of correlation, since the distribution of sample r's approaches the normal curve only as N becomes large, while the ratio $\frac{r\sqrt{N-m}}{\sqrt{1-r^2}}$ follows the t distribution whatever the size of the sample. The required value of t for various levels of significance and various degrees of freedom has been tabulated in convenient form, as in Appendix F. This table includes all values of n from 1 to 30; for larger values of n, we may use the last row of the t table, which is taken from the table of normal curve areas.

Analysis of variance. An alternative to the t test, which has some special advantages in connection with certain types of correlation to be treated in subsequent chapters, is the analysis of variance. The elements of this method were developed in Chapter XIII. The procedure is: (1) compute explained variance and unexplained variance based upon degrees of freedom; (2) compute the value of

$$z = \frac{1}{2}\log_e \frac{\bar{\sigma}_{y_c}^2}{\bar{\sigma}_{y_s}^2} = 1.15129 \log_{10} \frac{\bar{\sigma}_{y_c}^2}{\bar{\sigma}_{y_s}^2}, \text{ or } F = \frac{\bar{\sigma}_{y_c}^2}{\bar{\sigma}_{y_s}^2};$$

(3) determine whether the explained variance is significantly greater than the unexplained by reference to the z table (appendix G 1) or the F table (appendix G 2). The variances are obtained by dividing the sums of squared deviations by the appropriate degrees of freedom, as follows:

Explained: $\qquad \bar{\sigma}_{y_c}^2 = \dfrac{\Sigma y_C^2}{m-1} = \Sigma y_C^2.$

Unexplained: $\qquad \bar{\sigma}_{y_s}^2 = \dfrac{\Sigma y_S^2}{N-m} = \dfrac{\Sigma y_S^2}{N-2}.$

Total: $\qquad \bar{\sigma}_y^2 = \dfrac{\Sigma y^2}{N-1}.$

Applying this method to the common stock illustration, we obtain these results:

Source of variance	Variance
Explained .	$7{,}570.5 \div 1 = 7{,}570\ 5$
Unexplained .	$3{,}762.8 \div 62 = 60.690$
Total . .	$11{,}333.3 \div 63 = 179\ 894$

$$F = \frac{7{,}570\ 5}{60\ 690} = 124\ 74$$

Referring to Appendix G 2, we find that, when $n_1 = 1$ and $n_2 = 60$ (there is no entry for $n_2 = 62$), the .001 level of significance requires that $F = 11.972$. Since we have a value for F several times that size, we can unhesitatingly say that the explained variance is significantly greater than the unexplained, and therefore the correlation between the two variables is significant.

The Z transformation. Although the last two methods take care of small samples, they are applicable only for testing whether the coefficient is significantly different from zero. Since the sampling distribution of $\dfrac{r\sqrt{N-2}}{\sqrt{1-r^2}}$ becomes more and more skewed as r_P departs from zero, it is advisable to transform r into a measure, the sampling distribution of which is approximately normal, if we wish to test the divergence of a correlation coefficient from some hypothetical value other than zero or to test the difference between two correlation coefficients. This may be accomplished by means of the Z transformation described by Fisher.[20] Thus r may be transformed into Z by the following formula.

$$Z = \frac{1}{2}\left[\log_e(1+r) - \log_e(1-r)\right]$$

$$= \frac{1}{2}\log_e\frac{1+r}{1-r} = 1.15129\log_{10}\frac{1+r}{1-r}.$$

In the present instance

$$Z = 1.15129\log\frac{1.8173}{.1827} = 1.1486.$$

The standard error of Z is independent of r_P;

$$\sigma_Z = \frac{1}{\sqrt{N-m-1}}.$$

Thus

$$\sigma_Z = \frac{1}{\sqrt{61}} = .12804.$$

---

[20] R. A. Fisher, *Statistical Methods for Research Workers*, pp. 202–210, Oliver and Boyd, Edinburgh, 1938 (7th edition).

Although not quite so accurate as the t transformation for testing the hypothesis that r_P is zero, the Z transformation is substantially accurate. For the problem in hand,

$$\frac{Z}{\sigma_Z} = \frac{1.1486}{.12804} = 8.97.$$

Since the Z distribution is almost normal, the significance of Z may be determined by reference to the normal curve areas. Since 8.97 is beyond the table limits, Z (and therefore r) is unquestionably significant.

As stated above, the special province of Z is in testing the significance of the difference between the sample r and some known or hypothetical population value, or between two sample correlations. In order to do this, it is necessary also to transform the known or hypothetical r_P, or the other sample r, into Z.

Suppose, for instance, that we wish to test whether our r of $+.8173$ is significantly different from a hypothetical r_P of $+.7500$ When $r_P = +.7500$

$$Z_P = 1.15129 \log_{10} \frac{1\ 7500}{.2500} = .9730.$$

Remembering now that, when $r = +.8173$, $Z = 1.1486$, and referring to this value of Z as Z_1, we may compute

$$\frac{Z_1 - Z_P}{\sigma_{Z_1}} = \frac{1\ 1486 - .9730}{.12804} = \frac{.1756}{.12804} = 1\ 37.$$

Appendix E tells us that we may expect so large a difference from chance causes about 17 times in 100. Hence we must conclude that the difference is not significant. If, however, we are testing the significance of the difference between our r of $+\ 8173$ and another *sample* r of $+.7500$ ($Z_2 = .9730$) computed from 39 pairs of items, we must compute also

$$\sigma_{Z_2} = \frac{1}{\sqrt{39 - 2 - 1}} = \frac{1}{6} = .16667.$$

Then

$$\sigma_{Z_1 - Z_2} = \sqrt{(.12804)^2 + (.16667)^2} = .2102,$$

and

$$\frac{Z_1 - Z_2}{\sigma_{Z_1 - Z_2}} = \frac{.1756}{.2102} = .84.$$

Our table of normal areas tells us that a difference as large as the one obtained is to be expected (even if the two samples are drawn from the same population) about 40 times in 100. The difference between these two samples, therefore, is not significant.

As mentioned earlier, Z may be used to ascertain fiducial limits of r_P. The procedure consists in determining the fiducial limits of Z for the de-

sired level of significance and the proper degrees of freedom, and convert‑ing each of the two values of Z to r.

Correlation of Ranked Data

Sometimes statistical series are composed of items the exact magnitude of which cannot be ascertained but which are ranked according to size. Thus, in column 2 of Table 162, we have listed eight tennis players in order of their official ranking in 1936 by the United States Lawn Tennis Association. Because we wish to inquire whether tennis ability ran true to form in 1937, we have given their 1937 ranking order in column 3.

TABLE 162

Computation of Values for Correlation of Ranked Data: United States Lawn Tennis Association Rankings, 1936 and 1937

Player	ᵃ Ranking		Difference in rank [D = Col 2 − Col 3]		D^2
	1936	1937	+	−	
(1)	(2)	(3)	(4)	(5)	(6)
J Donald Budge	1	1	..	.	..
Frank A Parker . .	2	3	.	1	1
Bryan M Grant	3	4	.	1	1
Robert L. Riggs	4	2	2	..	4
John Van Ryn	5	8	..	3	9
Joseph R Hunt .. .	6	5	1	.	1
Harold Surface, Jr......	7	6	1		1
C. Gene Mako..... ..	8	7	1	..	1
Total	..		5	5	18

Source: United States Lawn Tennis Association as reported by tne New York *Times*

(As the table stands, we have the first eight of the twenty ranking players in the United States in 1936, who were also listed among the first twenty in 1937.)

Since the coefficient of correlation previously explained is not designed to deal with ranked data, we shall use *Spearman's rank correlation coefficient*, usually designated by the symbol ρ, the formula for which is

$$\rho = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)},$$

in which D refers to the difference in rank between paired items in the two series. (This coefficient must not be confused with coefficient of non-linear correlation, which customarily uses the same symbol.) In Table

162, it will be seen that the sum of the positive differences equals the sum of the negative differences, and thereby provides a check on the accuracy of the subtractions. Substituting the values in the formula, we have

$$\rho = 1 - \frac{6(18)}{8(64 - 1)} = +.786.$$

The formula gives the sign of the correlation, positive in this case. Whenever there is a tie in rank, the two or more positions should be split among the different items. Thus, had Riggs and Parker tied for second and third in 1937, each would have been ranked 2.5; while if Riggs, Parker, and Grant had tied for second, third, and fourth, each would have received a rank of 3.

This formula may also be used for ordinary data by converting the numerical data into ranks. For instance, we may rank American League baseball players according to their batting averages. A coefficient for eight such baseball averages, selected in the same manner as were the tennis data, yields a rank correlation coefficient of $-.143$, which seems to indicate that tennis form is more consistent than baseball batting form. One reason for using the rank method rather than the more exact method, even when actual values are available, is to save time. This saving is greatest when there are not very many items to be ranked. Since a correlation coefficient is not very reliable when the number of items is small, it may sometimes be desirable to make an estimate of the degree of association by use of the rougher and more quickly computed ρ.

The reason the rank method is not so accurate as the ordinary method is that all of the information concerning the data is not utilized. Thus the first differences of the values of the items in a series arranged in order of magnitude are almost never constant; usually these differences become smaller toward the middle of the array. If such first differences were constant, then r and ρ would give identical results. If the values, however, are distributed normally, there may be applied to ρ a correction which will give the same results that would be obtained directly by computing r.[21] These corrections always serve to increase the correlation; however, they are very small, in no case increasing the correlation by so much as .02. Furthermore, the correction is not always appropriate. In the present illustration we have only the upper tails of (possibly) normal distributions; if plotted, they would probably appear as reverse J distributions.

---

[21] Tables of corrected values of ρ are given in some textbooks. See, for instance, R. E. Chaddock, *Principles and Methods of Statistics*, p. 300 and Appendix E; Houghton Mifflin Company, Boston, 1925.

Correlation of Qualitative Distributions

Fortune Magazine printed a survey of public opinion in its October 1937 issue, and among the topics included was the popular viewpoint on the issue of third terms for Presidents of the United States. The following information is derived from a table on page 150 of that issue. If the rich

Attitude toward third term	Rich	Poor	Total	Per cent
Favorable	508	1,559	2,067	50 587
Unfavorable ..	905	1,114	2,019	49 413
Total . . .	1,413	2,673	4,086	100 000
Per cent...	34 581	65 419	100 000	..

and the poor were equally favorable to the principle of a third term, we should expect the following percentage distribution in the different cells:

Attitude	Rich	Poor
Favorable..	34.581 × 50 587 = 17.49	65 419 × 50.587 = 33 09
Unfavorable	34 581 × 49.413 = 17.09	65.419 × 49.413 = 32 33

These four percentages total 100; that is, 17.49 + 17.09 + 33.09 + 32.33 = 100.00. Applying these percentages to the total number of observations (4,086), we should expect the following distribution of occurrences:

Attitude	Rich	Poor	Total
Favorable. ....,...	715	1,352	2,067
Unfavorable	698	1,321	2,019
Total	1,413	2,673	4,086

We may now compare the observed with the expected frequencies, and compute χ^2 as in Table 163 (ordinarily only this table would be constructed; the others are purely expository).

$$\chi^2 = \Sigma \frac{(f - f_C)^2}{f_C} = 185.447.$$

There is only one degree of freedom for these data: if any one $f - f_C$ value is taken, the other $f - f_C$ values are determined by the requirement that the total number of rich, poor, favorable, and unfavorable be the same as the observed values. A practical method of computing the degrees of freedom lost is to add the number of columns to the number of rows and

subtract 1. The χ^2 table (appendix I) for the .001 level of significance requires that $\chi^2 = 10.827$. It is, therefore, almost inconceivable that we could obtain a χ^2 as high as 185.447 if attitude toward the question of the third term were not (in 1937) related to economic status. That hypothesis must therefore be discarded; the relationship between attitude and economic status must be held to be significant.

A further question is: How close is the relationship? This may be answered by computing the *coefficient of mean square contingency*.

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}} = \sqrt{\frac{185.447}{4,086 + 185.447}} = .21.$$

The χ^2 test and the coefficient of mean square contingency can be used when there are two or more categories for each variable.

One serious limitation to C is that its maximum value varies with the number of cells in the table, and therefore the values of C obtained from

TABLE 163

COMPUTATION OF χ^2 FOR ATTITUDE OF PERSONS ON THIRD TERMS FOR PRESIDENTS, BY ECONOMIC CLASSES

Attitude and economic class	Observed f	Expected f_c	Difference $f - f_c$	$(f - f_c)^2$	$\dfrac{(f - f_c)^2}{f_c}$
Rich.					
Favorable	508	715	−207	42,849	59.929
Unfavorable . .	905	698	207	42,849	61 388
Poor					
Favorable	1,559	1,352	207	42,849	31.693
Unfavorable	1,114	1,321	−207	42,849	32.437
Total	4,086	4,086	..	...	185.447

Source· Derived from data on p 150 of "Fortune Quarterly Survey. X," *Fortune*, Vol XVI, No. 4, October 1937

tables with different groupings are not comparable. The maximum value of C for a 2×2 table is .707; for a 5×5 table, .894; for a 10×10 table, .949; and the value approaches 1 as the number of classes is increased. The coefficient of mean square contingency (which has no sign) somewhat resembles the coefficient of correlation of grouped data, since the value of C approaches that of r as the number of classes is increased, provided certain conditions (regarding size of sample, normality of distribution, and arrangement of categories) are met.

Another method, known as *Sheppard's method of unlike signs*, may be used for preliminary investigations. It is given by the expression cos U

1.8°, where U is the percentage of cases of unlike sign. In the present instance,

$$U = \frac{508 + 1,114}{4,086} = 39.7 \text{ per cent,}$$

and the coefficient is

$$\cos (39.7) \ 1.8° = \cos 71 \ 46° = .32.$$

Appendix L gives a table of cosines. Whenever appropriate, a sign may be attached to C or to Sheppard's coefficient.

Selected References

B. H. Camp: *The Mathematical Part of Elementary Statistics*, Part I, Chapters VIII, IX, X, and Part II, Chapter V; D. C Heath and Co., Boston, 1934.

R. E. Chaddock: *Principles and Methods of Statistics*, Chapter XII; Houghton Mifflin Co , Boston, 1925. This approach treats r as the slope of the estimating line in units of σ.

F. E. Croxton and D. J. Cowden: *Practical Business Statistics*, Chapter XIX; Prentice-Hall, Inc , New York, 1934.

G. R. Davies and W. F. Crowder: *Methods of Statistical Analysis in the Social Sciences*, pages 226–250; John Wiley and Sons, New York, 1933. The approach is through comparison of diagonal deviations.

H. T. Davis and W. F. C. Nelson: *Elements of Statistics*, Chapter X; Principia Press, Bloomington, Indiana, 1935. A section on the properties of the correlation coefficient includes derivation of the rank correlation formula.

W. P. Elderton: *Frequency Curves and Correlation* (Third Edition), Chapters VII, VIII, IX; Cambrdge University Press, Cambridge, 1938.

W. P. Elderton and Ethel M. Elderton: *Primer of Statistics*, Chapter V, Adam and Charles Black, London, 1914. An exposition in which r is considered the slope of the estimating equation in units of standard deviations.

Mordecai Ezekiel: *Methods of Correlation Analysis*, Chapters I, II, III, IV, V, VII, VIII, IX; John Wiley and Sons, New York, 1930. This is the most comprehensive treatment of correlation available, and a reading of the entire book is recommended.

R. A. Fisher: *Statistical Methods for Research Workers* (Seventh Edition), pages 197–211, Chapter VII; Oliver and Boyd, Edinburgh, 1938. Pages 197–211 treat of the reliability of r, including the Z transformation. Chapter VII is an explanation of "intraclass" correlation.

F. C. Mills. *Statistical Methods Applied to Economics and Business* (Revised Edition), Chapter X; Henry Holt and Co., New York, 1938.

H. T. Rietz, Editor: *Handbook of Mathematical Statistics*, pages 132–138; Houghton Mifflin Co., Boston, 1924. A number of types of correlation are briefly described.

J. R. Rigglemann and I. N. Frisbee: *Business Statistics* (Second Edition), pages 244–256; McGraw-Hill Book Co., New York, 1938. A simple nonmathematical explanation of the meaning of correlation.

J. G. Smith: *Elementary Statistics*, Chapter XX and pages 388–398; Henry Holt and Co., New York, 1934. Chapter XX presents the product-moment ap-

proach very clearly, and also contains an elementary exposition of the bivariate normal frequency surface.

G. W. Snedecor: *Statistical Methods Applied to Experiments in Agriculture and Biology*, Chapters 6 and 7; Collegiate Press, Ames, Iowa, 1937. A discus, sion of co-variance will be found in Chapter 12

L. H. C. Tippett: *The Methods of Statistics* (Second Edition), Chapters VII and VIII; Williams and Norgate, Ltd., London, 1937. Chapter VII includes discussion of frequency surfaces and analysis of variance. Chapter VIII deals with sampling errors of correlation and regression constants. Scope of Chapter VIII is broader than included in this text.

Helen M. Walker: *Studies in the History of Statistical Method*, Chapter V; Williams and Wilkins, Baltimore, 1929.

A. E. Waugh: *Elements of Statistical Method*, Chapter IX; McGraw-Hill Book Co., New York, 1938.

G. U. Yule and M. G. Kendall: *An Introduction to the Theory of Statistics*, (Eleventh Edition), Chapters 5, 11, 12, 13; Charles Griffin and Co., London, 1937. Chapter 5 is a discussion of contingency tables and the coefficient of contingency. Chapters 11, 12, 13 are a thorough discussion of the theory of correlation, including simple correlation, contingency, rank correlation, grade correlation, tetrachoric r, and intra-class correlation.

CHAPTER XXIII

NON-LINEAR CORRELATION

The preceding chapter considered the simplest type of relationship between two variables: a constant amount of increase in the dependent variable associated with a unit increase in the independent variable. Not always, however, is the linear hypothesis satisfactory. Although it may be practical to estimate the height growth of forest trees from the increase in their breast-height diameter by a straight line equation, still the method has limitations. The estimating equation was found to be $Y_C = 1.045 + 1.677X$. But it seems unlikely that a tree which has not grown in diameter during 10 years would, nevertheless, have grown a foot in height. Thus, while the equation may be used to make estimates, it is probably unsatisfactory as the formulation of a scientific law. Any equation for these data which is sound theoretically should define a line which passes through the point $X = 0$, $Y = 0$. Even with such an equation, any estimate beyond the range of the data should be considered only as a hypothesis to be tested.

The social sciences abound in non-linear relationships. In the field of economics, for instance, demand curves are seldom straight lines. Again, the law of diminishing returns as stated by one well-known economist reads: "If additional equal quantities of a variable element are added to a fixed element, the additional output at first increases but eventually declines and finally becomes less than zero."[1] As thus stated, a mathematical formulation of the law would describe a curve with two bends in it— an equation with four constants—perhaps of the type $Y_C = a + bX + cX^2 + dX^3$.

Transforming Data to Linear Form

Chart 222 is a scatter diagram showing the relationship between production and real price of late cabbage in the United States during the years 1920–1935. "Real price" means price relative to the prices of other

---

[1] A. G. Black in *Economic Principles and Problems*, Walter E. Spahr, Editor, Farrar and Rinehart, New York, 1936 (3rd edition). p. 113.

commodities in general: in this instance, real prices are the nominal prices divided by index numbers of wholesale prices of all commodities, with the year 1926 taken as 100 per cent. The real price is thus an estimate of what cabbage would have sold for if the price of commodities in general had not changed. In this illustration each series has also been adjusted for trend, the trends selected being weighted moving averages which are thought to approximate the combined primary × secondary trends. Since the cyclical movements are not well-defined, the fluctuations in production
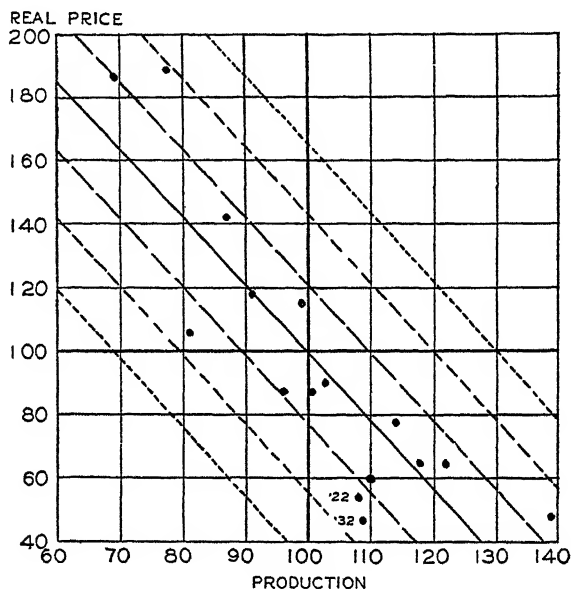
Chart 222. Production and Price of Late Cabbages in the United States, 1920-1935, and Zones of Scatter. Estimating equation: $Y_C = 313.3790 - 2.142386X$, shown by solid line. (Data of Table 164.)

are thus largely irregular fluctuations, and the price fluctuations are those which, by hypothesis, result from the variation in production. Although the coefficient of correlation is $-.862$, the estimating line does not seem to describe the relationship so well as might be desired. No doubt a more complex equation could be selected which would fit the data more closely, but it is always well to give preference to equation types with a small number of constants when so doing does not give a markedly poorer fit. The simpler the equation, the more reliable are the results. The reason for this is easy to understand.

Suppose a half dozen points are set at random upon a scatter diagram. If an equation type with six constants is selected, a curve can be fitted to

pass through all six points. The correlation thereby appears to be perfect when, as a matter of fact, the relationship between the two imaginary variables is entirely random. The paradox is explained when we realize that increasing the number of constants has an effect similar to reducing the size of the sample. A curve with six constants passing through six points is no more reliable than the mean of a sample of one item. In technical language, each time a constant is added, one degree of freedom is sacrificed, and thus, when a six-constant curve is fitted to six points, there are no degrees of freedom remaining. In this section, illustrations will be given of the transformation of data into such a form that a linear equation can reasonably be used. This practice has the desirable effect of reducing the constants to two, a and b in the equation type $Y_C = a + bX$.

The formulae that will be used will be similar to those used in simple correlation. It will be recalled that two normal equations were set up:

$$\text{I.} \quad \Sigma Y = Na + b\Sigma X.$$
$$\text{II.} \quad \Sigma XY = a\Sigma X + b\Sigma X^2.$$

From these the estimating equation $Y_C = a + bX$ was obtained. Using the constants of this equation, an expression $a\Sigma Y + b\Sigma XY = \Sigma Y_C^2$ was computed, and this expression was used in obtaining both σ_{y_s} and r by the formulae

$$\sigma_{y_s}^2 = \frac{\Sigma Y^2 - (a\Sigma Y + b\Sigma XY)}{N};$$

$$r^2 = \frac{(a\Sigma Y + b\Sigma XY) - \overline{Y}\Sigma Y}{\Sigma Y^2 - \overline{Y}\Sigma Y}.$$

Since $\Sigma Y_C^2 = a\Sigma Y + b\Sigma XY$, the above formulae for $\sigma_{y_s}^2$ and r^2 may be rendered simpler in appearance by writing them·

$$\sigma_{y_s}^2 = \frac{\Sigma Y^2 - \Sigma Y_C^2}{N}, \text{ or simply } \frac{\Sigma y_S^2}{N};$$

$$r^2 = \frac{\Sigma Y_C^2 - \overline{Y}\Sigma Y}{\Sigma Y^2 - \overline{Y}\Sigma Y}, \text{ or simply } \frac{\Sigma y_C^2}{\Sigma y^2}.$$

Writing the formulae thus may also help us to understand their meaning. Referring to the coefficient of determination r^2, we find that the numerator is the *explained variation* (which becomes the explained variance if divided by N), and that the denominator is the *total variation* (which becomes the total variance if divided by N); hence r^2 may be thought of as the ratio of the explained to the total variation (as well as the ratio of the explained to the total variance). The numerator of the fraction is made up of two parts: the explained sum of squares ΣY_C^2; and a correction factor $\overline{Y}\Sigma Y$, which is subtracted, leaving the numerator as

the explained sum of squared deviations, or simply the explained variation (see Appendix B, section XXII–1, equation 4). Likewise the denominator consists of the total sum of squares ΣY^2 and the same correction factor as before. The denominator is thus the sum of squared deviations, or simply total variation (see Appendix B, section XIII–2, part B–1). Since the correction factor is the same for both explained and total variation, when the explained variation is subtracted from the total variation, we obtain

$$(\Sigma Y^2 - \bar{Y}\Sigma Y) - (\Sigma Y_C^2 - \bar{Y}\Sigma Y) = \Sigma Y^2 - \Sigma Y_C^2,$$

which is the *unexplained variation,* the numerator of the formula for $\sigma_{y_s}^2$.

Use of logarithms. A scatter diagram of the cabbage data is shown as Chart 222. It is apparent that the relationship departs from linearity. We saw in Chapter XVI that a time series, the trend of which is concave upward sometimes becomes straight when plotted on semi-logarithmic paper, or when the logarithms are plotted on arithmetic paper. This is true also of data plotted in the form of a scatter diagram. Before actually transforming the data into logarithms, however, it is well to plot them first on paper which has been ruled logarithmically. If semi-logarithmic paper is used, we could alternately try plotting the dependent variable and the independent variable on the logarithmic scale. Paper is also available that is ruled logarithmically on both axes. Using this latter type of paper the cabbage data show a linear relationship in Chart 223A. This means that a constant percentage increase in real price seems to be associated with a constant percentage decrease in production.

Plotting on paper with logarithmic axes is equivalent to plotting the logarithms of the X values and of the Y values on arithmetic paper. This has been done in Chart 223B. Computation of the various measures of relationship involves procedures analogous to those already used. Table 164 gives the computation of values required. The formulae and their solution are below:

Equation type:

$$\log Y_C = \log a + b \log X,$$

which is the linear form of

$$Y_G = aX^b.$$

Normal equations:

I. $\quad \Sigma \log Y = N \log a + b\Sigma \log X;$
II. $\quad \Sigma(\log X \cdot \log Y) = \log a\Sigma \log X + b\Sigma (\log X)^2.$

These equations yield, in terms of logarithms, a straight estimating line from which the squares of the deviations of the logarithms are at a minimum. It is not, therefore, a least square fit to the original data, though the discrepancy is usually not large.
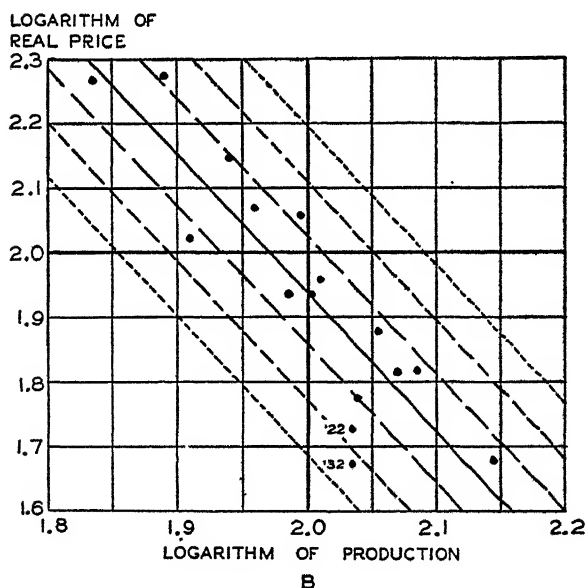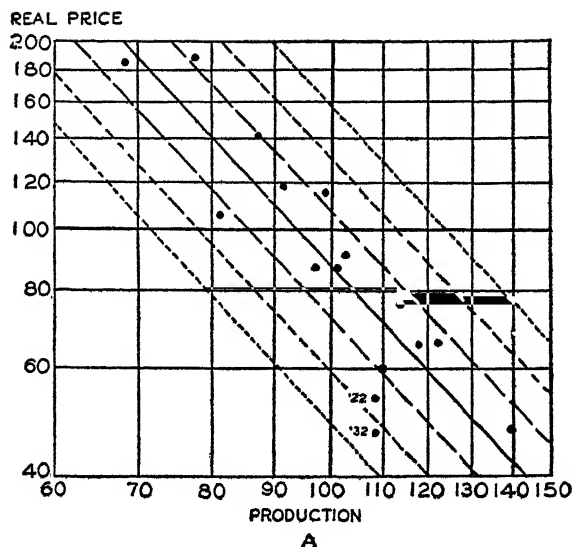
Chart 223. Production and Price of Late Cabbages and Zones of Scatter: A. Plotted on Logarithmic Paper, and B. Logarithms of Data Plotted on Arithmetic Paper. Estimating equation: Log $Y_c = 6.239331 - 2.149117 \log X$. (Data of Table 164.)

TABLE 164

COMPUTATION OF VALUES USED IN DETERMINING LOGARITHMIC MEASURES OF RELATIONSHIP BETWEEN PRODUCTION AND REAL PRICE OF LATE CABBAGE, 1920-1935

Year	Production (per cent of normal) X	Real price (per cent of normal) Y	log X	log Y	(log X)(log Y)	(log X)²	(log Y)²
1920	139.3	48.0	2 143951	1.681241	3 604498	4 596526	2 826571
1921	68 6	186.3	1 836324	2 270213	4 168847	3 372086	5 153867
1922	108 3	53 6	2 034628	1 729165	3 518208	4 139711	2 990012
1923	81 0	106 3	1 908485	2.026533	3 867608	3.642315	4 106836
1924	113.7	77 5	2 055760	1.889302	3 883951	4 226149	3 569462
1925	99.1	115.3	1 996074	2.061829	4 115563	3 984311	4 251139
1926	102 5	89.8	2.010724	1 953276	3 927499	4 043011	3 815287
1927	118 0	65.2	2.071882	1.814248	3 758908	4 292635	3 291496
1928	87 3	141.7	1.941014	2.151370	4.175839	3 767535	4 628393
1929	91 0	118.1	1.959041	2 072250	4 059623	3 837842	4 294220
1930	100.6	86 8	2 002598	1 938520	3 882076	4 010399	3 757860
1931	96.2	86 8	1.983175	1 938520	3 844424	3 932983	3 757860
1932	108 6	47.4	2.035830	1.675778	3.411599	4 144604	2 808232
1933	77.5	188.4	1 889302	2 275081	4 298315	3 569462	5 175994
1934	121.7	64.6	2 085291	1 810233	3 774863	4 348439	3 276944
1935	110.1	60 1	2 041787	1 778874	3.632082	4 168894	3 164393
Total			31 995866	31 066433	61 923903	64 076962	60 868566

Source: Data on production and price of late cabbage are from United States Department of Agriculture, *Agricultural Outlook Charts, Potatoes and Truck Crops, 1938*, p. 18. Deflator is United States Bureau of Labor Statistics Index of Wholesale Commodity Prices, All Commodities. Trends are 5-year weighted moving averages.

Substituting values from Table 164 in these normal equations, we have:

I. $31.066433 = 16 \log a + 31.995866b$;
II. $61.923903 = 31.995866 \log a + 64.076962b$.

These give the estimating equation

$$\log Y_C = 6.239331 - 2.149117 \log X;$$

or in terms of the original data

$$Y_C = 1,735,128 X^{-2.149117}.$$

The explained sums of squares are:

$$\begin{aligned}
\Sigma (\log Y_C)^2 &= \log a \Sigma \log Y + b\Sigma (\log X \cdot \log Y) \\
&= (6.239331)(31.066433) - (2.149117)(61.923903) \\
&= 60.752045.
\end{aligned}$$

To obtain the standard error of estimate, we compute as follows:

$$\sigma^2_{\log y_s} = \frac{\Sigma (\log y_s)^2}{N} = \frac{\Sigma (\log Y)^2 - \Sigma (\log Y_C)^2}{N}$$

$$= \frac{60.868566 - 60.752045}{16} = .0072826,$$

$$\sigma_{\log y_s} = .085340.$$

We may now proceed to find the zones of scatter, which are shown in Charts 223 and 224. $X = 100$ per cent will be used as the point of reference.

If $X = 100$,
$$\log X = 2.000000,$$

and substituting in the equation

$$\begin{aligned}
\log Y_C &= 6.239331 - 2.149117 \log X, \\
\log Y_C &= 1.941097, \\
Y_C &= 87.32 \text{ per cent.}
\end{aligned}$$

The values of $\log Y_C$ and $\sigma_{\log y_s}$ must be added before the anti-logarithm is obtained; thus

$$\begin{aligned}
\log Y_C + \sigma_{\log y_s} &= 1.941097 + .085340 = 2.02644. \\
\log Y_C - \sigma_{\log y_s} &= 1.941097 - .085340 = 1.85576.
\end{aligned}$$

Looking up the anti-logs of these values in Appendix P, we find:

$$\begin{aligned}
\text{antilog } (\log Y_C + \sigma_{\log y_s}) &= 106.28 \text{ per cent}; \\
\text{antilog } (\log Y_C - \sigma_{\log y_s}) &= 71.74 \text{ per cent.}
\end{aligned}$$

Similarly:

$$\text{antilog } (\log Y_C + 2\sigma_{\log y_s}) = 129.35 \text{ per cent;}$$
$$\text{antilog } (\log Y_C - 2\sigma_{\log y_s}) = 58.94 \text{ per cent.}$$

$$\text{antilog } (\log Y_C + 3\sigma_{\log y_s}) = 157.44 \text{ per cent;}$$
$$\text{antilog } (\log Y_C - 3\sigma_{\log y_s}) = 48.43 \text{ per cent.}$$

In a similar manner zones of scatter can be obtained for other values of X. We can also, however, express the standard error of estimate in
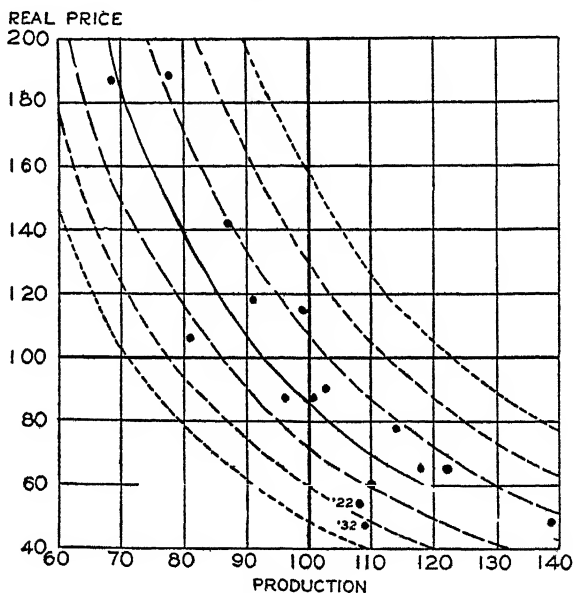
Chart 224. Production and Price of Late Cabbage and Zones of Scatter. Estimating equation: Log $Y_c = 6.239331 - 2.149117 \log X$. (Data of Table 164.)

the form of a scatter ratio by obtaining the anti-log of $\sigma_{\log y_s}$. Since $\sigma_{\log y_s}$ = .085340, plus one scatter ratio is 1.2171; and since $-\sigma_{\log y_s} = -.085340$ = 9.914660 − 10, minus one scatter ratio = .8216 (the reciprocal of 1.2171). These results indicate that, regardless of the value of Y_C,

$$\text{antilog } (\log Y_C + \sigma_{\log y_s}) = 1.2171 Y_C;$$
$$\text{antilog } (\log Y_C - \sigma_{\log y_s}) = .82160 Y_C.$$

When $Y_C = 87.32$, as in the illustration above,

$$\text{antilog } (\log Y_C + \sigma_{\log y_s}) = 1.2171(87.32) = 106.28 \text{ per cent;}$$
$$\text{antilog } (\log Y_C - \sigma_{\log y_s}) = .82160(87.32) = 71.74 \text{ per cent.}$$

These are the same values that were obtained above. The ratio used in obtaining antilog $(\log Y_C + 2\sigma_{\log y_s})$ is the antilog of $2\sigma_{\log y_s}$; or, the ratio may be obtained directly by squaring one scatter ratio. The other ratios may be similarly obtained.

The coefficient of curvilinear correlation is usually referred to as the *index of correlation*, and may be designated by the symbol ρ to distinguish it from the ordinary coefficient of correlation; ρ^2 may be called the *index of determination*. No positive or negative sign is set before ρ, though in the case of the cabbage data it might seem logical to consider the sign negative. However, for certain non-linear relationships, the slope is positive in some parts of the curve and negative in others. Illustrations of such curves will be found later in this chapter.

$\rho_{\log Y \log X}$ is easily obtained by substituting in the formula:

$$\rho^2_{\log Y \log X} = \frac{\Sigma(\log y_C)^2}{\Sigma(\log y)^2} = \frac{\Sigma(\log Y_C)^2 - (\overline{\log Y})\Sigma\log Y}{\Sigma(\log Y)^2 - (\overline{\log Y})\Sigma\log Y}.$$

where $(\overline{\log Y})$ is the mean of the log Y values.

$$\rho^2_{\log Y \log X} = \frac{60\ 752045 - (1.941652)(31.066433)}{60.868566 - (1.941652)(31.066433)} = \frac{.431843}{548364},$$

$$= .7875.$$

$$\rho_{\log Y \log X} = .887.$$

Using the logarithms of the X and Y observations has increased the correlation from $r = -.862$ to $\rho = .887$.

Of course, $\rho^2_{\log Y \log X}$ tells us the proportion of variation in (or variance of) the logarithms of the Y values that has been explained by reference to the logarithms of the X values. It is the ratio of the variation in the computed log Y values to the variation in the actual log Y values.

Also, $\rho_{\log Y \log X}$ can be obtained directly by the following expression, which parallels the formula used for r:

$$\rho_{\log Y \log X} = \frac{N\Sigma \log X \log Y - (\Sigma \log X)(\Sigma \log Y)}{\sqrt{[N\Sigma(\log X)^2 - (\Sigma \log X)^2][N\Sigma(\log Y)^2 - (\Sigma \log Y)^2]}}$$

$$= \frac{16(61.923903) - (31.995866)(31.066433)}{\sqrt{[16(64\ 076962) - (31.995866)^2][16(60.868566 - (31.066433)^2]}}$$

$$= .887.$$

This type of formula cannot be used when there are more than two constants in the estimating equation.

Use of reciprocals. Since the price of cabbage decreases as production

increases, it is not unreasonable to hypothesize that the relationship is reciprocal—that it may be described by an equation of the type[2]

$$\frac{1}{Y_c} = a + bX.$$

Reciprocals of the Y values were obtained from Appendix O. These have been plotted against the X values in Chart 225. The results seem to discredit the reciprocal hypothesis. The values for 1922 and 1932, which were slightly too low, in Chart 222, now seem much too high. Although the relationship seems linear, except for these two observations.

TABLE 165

COMPUTATION OF VALUES USED IN DETERMINING RECIPROCAL MEASURES OF RELATIONSHIP BETWEEN PRODUCTION (X) AND REAL PRICE (Y) OF LATE CABBAGE, 1920–1935

(For convenience in computation the X and Y variables, which are ratios to trend, have been considered as decimals rather than percentages.)

Year	X	$\frac{1}{Y}$	$X\left(\frac{1}{Y}\right)$	X^2	$\left(\frac{1}{Y}\right)^2$
1920	1 393	2 083333	2 902083	1.940449	4 340276
1921	686	536769	368224	.470596	288121
1922	1 083	1 865672	2 020523	1 172889	3 480732
1923	810	940734	761995	656100	.884980
1924	1.137	1 290323	1 467097	1 292769	1 664933
1925	.991	867303	859497	.982081	752214
1926	1 025	1 113586	1 141426	1.050625	1.240074
1927	1.180	1 533742	1 809816	1 392400	2.352365
1928	873	705716	616090	762129	.498035
1929	910	846740	770533	828100	.716969
1930	1.006	1 152074	1 158986	1.012036	1 327275
1931	962	1 152074	1.108295	.925444	1 327275
1932	1.086	2.109705	2.291140	1 179396	4 450855
1933	.775	.530786	411359	600625	.281734
1934	1 217	1 547988	1.883901	1 481089	2 396267
1935	1 101	1 663894	1 841931	1.212201	2.768543
Total	16 235	19 940439	21.412896	16.958929	28 770648

Source of data: Table 164

it is not apparent that the correlation is higher than for the arithmetic relationship, and it seems lower than for the logarithmic relationship.

[2] Alternately, the type $Y_c = a + b\frac{1}{X}$ could be used: if the Y values of this illustration are plotted against the reciprocals of the X values, it will be noticed that the relationship between the variables is linear, and is much closer than that shown in Chart 225. Nevertheless, the equation type $\frac{1}{Y_c} = a + bX$ was chosen for purposes of exposition, since this equation involves a problem in connection with the standard error of estimate not encountered with the former type.

It is possible to correlate the X values with the $\frac{1}{Y}$ values in a fashion precisely similar to that with which the reader is familiar. The normal equations are of the following type:

$$\text{I.} \quad \Sigma\left(\frac{1}{Y}\right) = Na + b\Sigma X;$$

$$\text{II.} \quad \Sigma\left(X\frac{1}{Y}\right) = a\Sigma X + b\Sigma X^2.$$
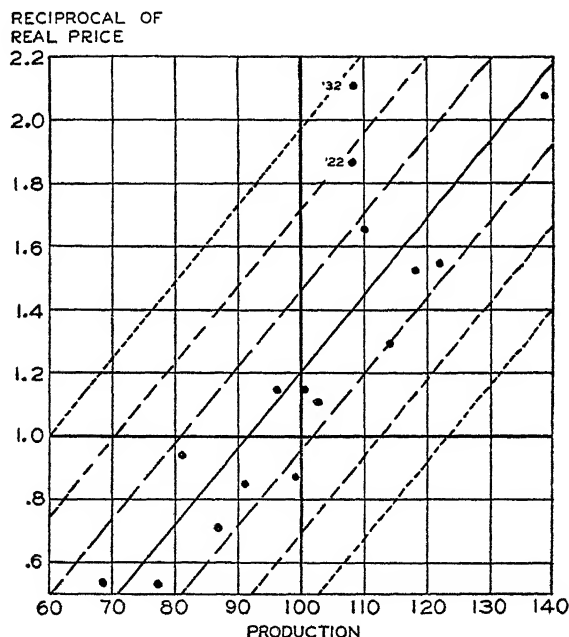
Chart 225.   Production and Reciprocals of Price of Late Cabbage and Zones of Scatter.
Estimating equation: $\dfrac{1}{Y_c} = -1.219147 + 2.429738X$.   (Data of Table 165.)

The necessary computations are shown in Table 165. Substituting the values obtained, we have:

$$\text{I.} \quad 19.940439 = 16a + 16{,}235b;$$

$$\text{II.} \quad 21.412896 = 16.235a + 16.9589296.$$

The estimating equation is found to be

$$\frac{1}{Y_c} = -1.219147 + 2.429738X.$$

Continuing in the usual fashion, we have for the explained sums of squares:

$$\Sigma\left(\frac{1}{Y_C}\right)^2 = a\Sigma\left(\frac{1}{Y}\right) + b\Sigma X\left(\frac{1}{Y}\right)$$

$$= (-1.219147)(19.940439) + (2.429738)(21.412896)$$

$$= 27\ 717401.$$

We now obtain the standard error of estimate as follows:

$$\sigma^2_{\frac{1}{y_s}} = \frac{\Sigma\left(\frac{1}{y_s}\right)^2}{N} = \frac{\Sigma\left(\frac{1}{Y}\right)^2 - \Sigma\left(\frac{1}{Y_C}\right)^2}{N}.$$

$$= \frac{28.770648 - 27.717401}{16} = .065828.$$

$$\sigma_{\frac{1}{y_s}} = .2566.$$

Chart 225, on which are plotted the reciprocals of the Y values, shows the estimating line and zones of $\pm\sigma_{\frac{1}{y_s}}$, $2\sigma_{\frac{1}{y_s}}$, $3\sigma_{\frac{1}{y_s}}$, while Chart 226 is the same except that natural numbers are plotted along both axes. A word of explanation is advisable concerning the method of obtaining the zones of scatter in Chart 226. For illustrative purposes we shall find the Y_C value necessary for plotting when production is 100 per cent. The Y_C values corresponding to other X values are obtained in a similar fashion.

If $X = 1.00$, substituting in the estimating equation

$$\frac{1}{Y_C} = -1.219147 + 2.429738X,$$

we find that

$$\frac{1}{Y_C} = 1.210591, \text{ and } Y_C = 82.60 \text{ per cent.}$$

$$\frac{1}{Y_C} - \sigma_{\frac{1}{y_s}} = 1.210591 - .2566 = .953991.$$

$$\frac{1}{Y_C} + \sigma_{\frac{1}{y_s}} = 1.210591 + .2566 = 1.467191.$$

Looking up the reciprocals of these values in Appendix O, we find

$$\text{reciprocal}\left(\frac{1}{Y_C} - \sigma_{\frac{1}{y_s}}\right) = 104.82 \text{ per cent, and}$$

$$\text{reciprocal}\left(\frac{1}{Y_C} + \sigma_{\frac{1}{y_s}}\right) = 68.16 \text{ per cent.}$$

It should be observed that the reciprocal values are combined before the final result is obtained.

Similarly: reciprocal $\left(\dfrac{1}{Y_C} - 2\sigma_{\frac{1}{y_s}}\right) = 143.39$ per cent, and

reciprocal $\left(\dfrac{1}{Y_C} + 2\sigma_{\frac{1}{y_s}}\right) = 58.01$ per cent;

reciprocal $\left(\dfrac{1}{Y_C} - 3\sigma_{\frac{1}{y_s}}\right) = 226.86$ per cent, and

reciprocal $\left(\dfrac{1}{Y_C} + 3\sigma_{\frac{1}{y_s}}\right) = 50.50$ per cent.

The zones of scatter for other values of X may be obtained in a similar fashion. The width of each zone will differ for each value of X, as can be seen by an inspection of Chart 226.
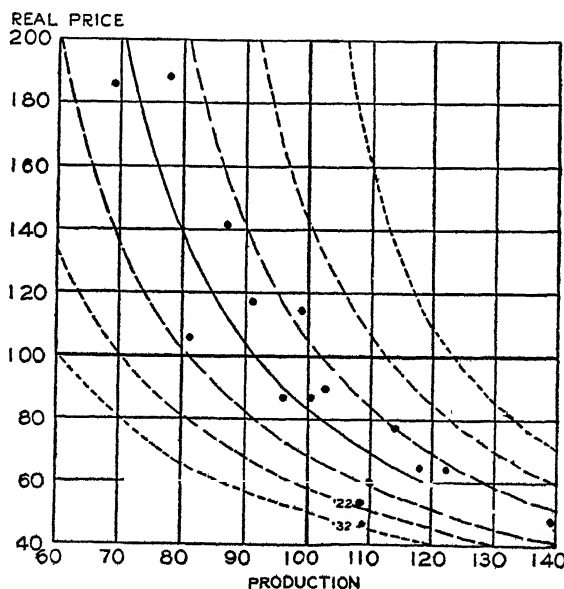
Chart 226. Production and Price of Late Cabbage and Zones of Scatter. Estimating equation: $\dfrac{1}{Y_C} = -1.219147 + 2.429738X$. (Data of Table 165.)

Just as r^2 may be thought of as the proportion of total variation in the original data that has been explained, so $\rho^2_{\frac{1}{Y}X}$ is the proportion of total variation in the reciprocals of the original data that has been explained. That is, it is the ratio of the variation in the computed reciprocals to the variation in the reciprocals of the original data. Thus

$$\rho^2_{\frac{1}{Y}X} = \frac{\Sigma\left(\dfrac{1}{y_c}\right)^2}{\Sigma\left(\dfrac{1}{y}\right)^2} = \frac{\Sigma\left(\dfrac{1}{Y_c}\right)^2 - \left(\overline{\dfrac{1}{Y}}\right)\Sigma\left(\dfrac{1}{Y}\right)}{\Sigma\left(\dfrac{1}{Y}\right)^2 - \left(\overline{\dfrac{1}{Y}}\right)\Sigma\left(\dfrac{1}{Y}\right)},$$

where $\left(\overline{\dfrac{1}{Y}}\right)$ is the mean of the $\dfrac{1}{Y}$ values.

$$\rho^2_{\frac{1}{Y}X} = \frac{27.717401 - 24.851310}{28.770648 - 24.851310}$$

$$= \frac{2.866091}{3.919338} = .7313.$$

$$\rho_{\frac{1}{Y}X} = .855.$$

As stated before, ρ has no sign. In the present instance, although the line of relationship is a linear fit in terms of reciprocals of Y and has a positive slope in terms of reciprocals (see Chart 225), yet, when the computed values are re-converted into the original units, the slope becomes negative.

When there are only two constants in the estimating equation, $\rho_{\frac{1}{Y}X}$ can be computed directly as usual:

$$\rho_{\frac{1}{Y}X} = \frac{N\Sigma\left(X\frac{1}{Y}\right) - \Sigma X\Sigma\left(\frac{1}{Y}\right)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2]\left[N\Sigma\left(\frac{1}{Y}\right)^2 - \left(\Sigma\frac{1}{Y}\right)^2\right]}}$$

$$= \frac{16(21.412896) - (16\ 235)(19.940439)}{\sqrt{[16(16.958929) - (16.235)^2][16(28.770648) - (19.940439)^2]}}$$

$$= .855.$$

The choice between the different methods of expressing relationship is not always clear. In this instance the coefficient of correlation is highest when the relationship is assumed to be logarithmic ($\rho = .887$), and smallest when it is assumed to be reciprocal ($\rho = .855$). However, the difference is not very large, and for purposes of estimation other factors must be taken into consideration.

By which method is the error of estimate reduced to smallest magnitude in absolute terms? As can be seen from Table 166, the range of error to be expected for 68.27 per cent of the items is 43.88 per cent for each value of X, by the arithmetic method. By the reciprocal method, the error is smallest for high values of X, but by far the largest for small values of X. When X is 60, the range is infinity. Another consideration is the distribution of the scatter around the different estimating lines. The distribution appears to be more nearly normal around the line of section B of Chart 223 than is the case with Chart 222 or Chart 225. Probably the logarithmic curve is the best of the three under consideration.

All three methods are consistent with the law of demand commonly set down by economists.   The logarithmic method also involves the assumption that the flexibility of price[3] is the same, regardless of market supply.

Curves with More than Two Constants

Second degree curve.   It is well known that the per capita expense of city administration increases with the size of the city.   For instance, in a large city many policemen are required to regulate traffic.   Congested

TABLE 166

Range of Standard Error Involved in Three Different Assumptions Concern- ing Relationship of Production and Real Price of Late Cabbage, 1920–1935

(Per cent)

X	Linear	Logarithmic	Reciprocal
		Y_c	
60	184 84	261.74	418 94
100	99 14	87.32	82.60
140	13.44	42.37	45.82
		$Y_c \pm \sigma_{y_s}$	
60	162 90–206 78	215.05–318.58	201.90–∞
100	77.20–121 08	71 74–106.28	68 16–104.82
140	−8 50–35 38	34.81–51 57	41 00–51.92
		Range from $-\sigma_{y_s}$ to $+\sigma_{y_s}$	
60	43 88	103.53	∞
100	43 88	34.54	36 66
140	43 88	16.76	10.92

Source· Derived from Tables 164 and 165

areas are a breeding ground for criminals, and the opportunity for crime is more prevalent than in rural areas.   As an illustration of the tendency of police expense per capita to increase with population of city, we use those cities between 50,000 and 300,000 population in 7 selected mid- western states.   It was necessary to choose states which as a group were

---

[3] The equation $Y_c = 1,735,128X^{-2.149117}$ indicates that the flexibility of price is −2.149117.   For a general method of determining flexibility of price at any value of X, see Appendix B, section XXIII–1.

fairly homogeneous; otherwise the tendency would be obscured. Also, it was deemed advisable to omit certain places (for example, Cicero, Illinois) which are in such close proximity to much larger cities that their police problem is closely tied up with the larger places. Only 17 cities are included in our sample. The results are therefore not very reliable, but the smallness of the sample facilitates the illustration of the application of the method.

From Chart 227 it is obvious that the relationship between the variables

PER CAPITA EXPENSE, DOLLARS

POPULATION IN THOUSANDS
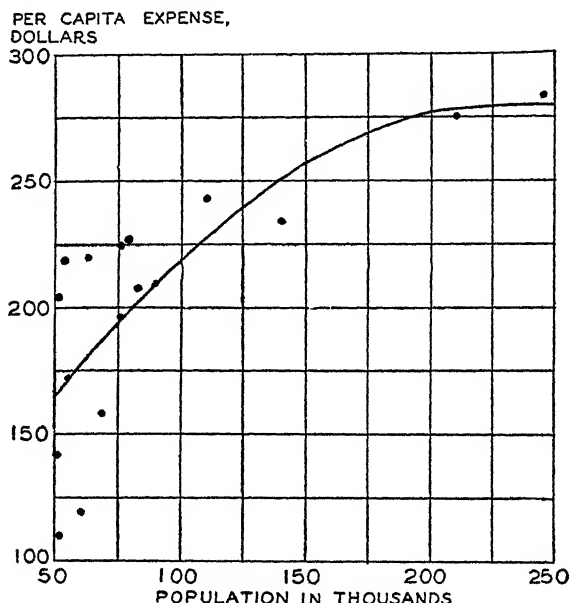
Chart 227. Population and Per Capita Police Department Expense of 17 Mid-Western Cities, 1925. (Group III and Group IV Cities of Wisconsin, Minnesota, Illinois, Iowa, Kansas, Nebraska, and Missouri. These cities range in population between 50,000 and 300,000. Kansas City, Kansas; East St. Louis, Illinois; Cicero, Illinois; and Oak Park, Illinois, are omitted, since they are part of metropolitan areas of much larger cities. Data from Bureau of the Census, *Financial Statistics of Cities,* 1925)

is curvilinear and that if an equation of the type $Y_C = a + bX + cX^2$ is used, b will be positive and (since the increment of increase is decreasing) c will be negative. The normal equations for a curve of this type are three, since there are three constants:

$$\text{I.} \quad \Sigma Y = Na + b\Sigma X + c\Sigma X^2;$$
$$\text{II.} \quad \Sigma XY = a\Sigma X + b\Sigma X^2 + c\Sigma X^3;$$
$$\text{III.} \quad \Sigma X^2Y = a\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4.$$

The computation of values required for these equations is given in Table 167.

Perhaps the simplest way[4] to solve these three equations simultaneously is:

(1) Solve equations I and II simultaneously so as to eliminate a, and thus leave a resulting equation A.

I. $34.55 = 17a + 1,564.9b + 195,852.89c.$
II. $3,504.60 = 1,564.9a + 195,852.89b + 32,660,157.11c.$

Multiplying I by 920.52941:

I. $3,180.429 = 1,564.9a + 144,053.65b + 18,028,035c$
II. $3,504.600 = 1,564.9a + 195,852.89b + 32,660,157c$
A. $\overline{324.171 = \qquad\qquad 51,799.24b + 14,631,322c.}$

(2) Similarly, solve equations II and III simultaneously so as to eliminate a. Call the resulting equation B.

B. $46,531.23 = 8,148,461b + 2,430,267,000c.$

(3) Solve equations A and B simultaneously so as to eliminate b. Thus, multiplying A by 157.30850:

A. $50,994.85 = 8,148,461b + 2,301,631,000c$
B. $\underline{46,531.23 = 8,148,461b + 2,430,267,000c}$
      $4,463.62 = \qquad\qquad\quad - 128,636,000c$
      $c = -.0000346996.$

(4) Substitute the value of c in either equation A or equation B.

B. $46,531.23 = 8,148,461b + (2,430,267,000)(-.0000346996)$
      $b = .0160595.$

---

[4] Simultaneous solution of three equations may be avoided by the following procedure. Set up normal equations I and II as follows:

$$\text{I.} \quad a = \frac{\Sigma Y - b\Sigma X - c\Sigma X^2}{N};$$

$$\text{II.} \quad a = \frac{\Sigma XY - b\Sigma X^2 - c\Sigma X^3}{\Sigma X}.$$

Then substitute these expressions in equations II and III respectively:

$$\text{II'.} \quad \Sigma XY = \left(\frac{\Sigma Y - b\Sigma X - c\Sigma X^2}{N}\right)\Sigma X + b\Sigma X^2 + c\Sigma X^3;$$

$$\text{III'.} \quad \Sigma X^2 Y = \left(\frac{\Sigma XY - b\Sigma X^2 - c\Sigma X^3}{\Sigma X}\right)\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4.$$

These two equations are then solved simultaneously for b and c, after which a is obtained by substitution of b and c in one of the normal equations.

TABLE 167

COMPUTATION OF VALUES REQUIRED FOR MEASUREMENT OF RELATIONSHIP BETWEEN POPULATION AND PER CAPITA POLICE DEPARTMENT
EXPENSE OF 17 MID-WESTERN CITIES, 1925

City	Population (thousands) X	Per capita police department expense Y	X²	X³	X⁴	XY	X²Y	Y²
St. Paul	246.0	$ 2.84	60,516.00	14,886,936.00	3,662,186,256	696 640	171,865.44	8 0656
Omaha	211.8	2.76	44,859.24	9,501,187.03	2,012,351,413	584 568	123,811.50	7 6176
Des Moines	140.7	2.36	19,796.49	2,785,366.14	391,901,016	332 052	46,719.72	5 5696
Duluth	110.5	2.44	12,210.25	1,349,232.62	149,090,205	269 620	29,793.01	5 9536
Wichita	89.4	2.10	7,992.36	714,516.98	63,877,818	187 740	16,783.96	4 4100
Peoria	81.6	2.08	6,658.56	543,338.50	44,336,421	169 728	13,849.80	4 3264
St Joseph	78.3	2.27	6,130.89	480,048.69	37,587,812	177 741	13,917.12	5 1529
Rockford	76.5	1.97	5,852.25	447,697.12	34,248,830	150 705	11,528.93	3 8809
Sioux City	76.2	2.25	5,806.44	442,450.73	33,714,745	171 450	13,064.49	5 0625
Racine	67.7	1.58	4,583.29	310,288.73	21,006,547	106 966	7,241.60	2 4964
Springfield	62.8	2.21	3,943.84	247,673.15	15,553,874	138 788	8,715.89	4 8841
Lincoln	60.6	1.18	3,672.36	222,545.02	13,486,228	71 508	4,333.38	1 3924
Topeka	55.7	1.73	3,102.49	172,808.69	9,625,111	96 361	5,367.31	2 9929
Decatur	53.2	1.11	2,830.24	150,568.77	8,010,258	59 052	3,141.57	1 2321
Davenport	52.7	2.19	2,777.29	146,363.18	7,713,310	115 413	6,082.27	4 7961
Kenosha	50.9	2.04	2,590.81	131,872.23	6,712,296	103 836	5,285.25	4 1616
Cedar Rapids	50.3	1.44	2,530.09	127,263.53	6,401,355	72 432	3,643.33	2 0736
Total	1,564.9	34.55	195,852.89	32,660,157.11	6,517,803,858	3,504 600	485,144.57	74 0683

Source: Based on data in United States Bureau of the Census, *Financial Statistics of Cities*, 1925, pp 86–87 and 270–274

Normal equations:

$$\bar{Y} = \frac{34.55}{17} = 2.03235.$$

$$\text{I.} \qquad 34.55 = 17a + 1,564.9b + 195,852.89c.$$

$$\text{II.} \qquad 3,504.600 = 1,564.9a + 195,852.89b + 32,660,157.11c.$$

$$\text{III.} \qquad 485,144.57 = 195,852.89a + 32,660,157.11b + 6,517,803,858c.$$

Estimating equation: $Y_C = .953795 + .0160595X - .0000346996X^2.$

Explained sums of square:
$$\Sigma Y_c^2 = a\Sigma Y + b\Sigma XY + c\Sigma X^2Y$$
$$= .953795(34.55) + .0160595(3504\ 600) - .0000346(485,144.57)$$
$$= 72.4014$$

(5)   Substitute the values of b and c in equation I, II, or III.

I. $34.55 = 17a + (1,564.9)(.0160595) + (195,852.89)(-.0000346996)$
      $a = .953795.$

(6)   Check accuracy by substituting values of a, b, and c in either of the two remaining normal equations.

II. $3,504.60 = (1,564.9)(.953795) + (195,852.89)(.0160595)$
      $$+ (32,660,157)(-.0000346996)$$
      $$= 3,504.599 = 3,504.60.$$

(7)   The estimating equation is

$$Y_C = .953795 + .0160595X - .0000346996X^2.$$

It should be noticed that six digits are included in the values of a, b, and c, though perhaps only five are significant for a. This necessitates more than six digits in the various equations necessary to obtain these values. The reason this is true is that the various multiplications required multiply the inaccuracies inherent in rounding. In general, however, it is better to show too many digits than too few. As the computations proceed toward their final conclusion, figures that lose their significance may be dropped.

From the estimating equation, the desired Y_C values may be computed as shown below. Values within the range of data only have been included,

X	X^2	$a + bX$	cX^2	Y $(a + bX + cX^2)$
50	2,500	1.756	− .087	$1 67
100	10,000	2.559	− 347	2.21
150	22,500	3.362	− 781	2 58
200	40,000	4 166	−1.388	2 78
250	62,500	4.969	−2 169	2.80

since in themselves the original observations afford no evidence beyond their range. Furthermore, in this instance it seems illogical even to hypothesize that the equation will be useful if extended far in either direction. Notice that if unduly extended the equation implies that a town without population would spend $.95 per capita on police, and that per capita expense would eventually decline, and even become negative, with increased size of city.

The formula for $\sigma_{v_s}^2$ is of the same type that was used in linear correlation:

$$\sigma_{v_s}^2 = \frac{\Sigma y_s^2}{N} = \frac{\Sigma Y^2 - \Sigma Y_C^2}{N} = \frac{\Sigma Y^2 - (a\Sigma Y + b\Sigma XY + c\Sigma X^2 Y)}{N}.$$

Proof that $\Sigma Y_C^2 = a\Sigma Y + b\Sigma XY + c\Sigma X^2Y$ is similar to that shown for equation 3 in Appendix B, section XXIII–1.

From Table 167 we see that $\Sigma Y_C^2 = 72.4014$; hence

$$\sigma_{y_s}^2 = \frac{74.0683 - 72.4014}{17} = .0985,$$

$$\sigma_{y_s} = \$.313.$$

We may now proceed to find ρ in the usual fashion.

$$\rho^2 = \frac{\Sigma y_C^2}{\Sigma y^2} = \frac{\Sigma Y_C^2 - \overline{Y}\Sigma Y}{\Sigma Y^2 - \overline{Y}\Sigma Y}$$

$$= \frac{72.4014 - (2.03235)(34.55)}{74.0683 - (2.03235)(34.55)} = \frac{2.18371}{3.85061} = .5671.$$

$$\rho = .753.$$

As before, ρ has no sign.

Test of fitness of equation type.   Although the reliability of a curvilinear correlation coefficient will be considered in the final section of this chapter, it is worth while to consider at this point whether the increase in the correlation by the introduction of an additional constant in the estimating equation is a significant increase.

Now it has been found that

$$\rho^2 = \frac{2.18371}{3.85061} = .5671; \text{ and } \rho = .753.$$

But use of the straight line equation gives these results:

$$r^2 = \frac{2.02893}{3.85061} = .5269; \text{ and } r = +.726.$$

(In the above expression, $2.02893 = \Sigma y_C^2$ for a linear equation the computation of which is not shown.   It may be obtained from the data given in Table 167.)   We may discover whether ρ is significantly higher than r by application of the analysis of variance technique which was outlined in Chapter XIII.

We may summarize our results from using a straight line estimating equation as follows:

Source of variation	Amount of variation	Degrees of freedom	Variance
Explained by straight line.	2 02893	1	2 02893
Unexplained by straight line..	1.82168	15	.12145
Total.................	3.85061	16	0.24066

From this summary we see that the squared residuals from the straight line total 1.82168. We may now inquire how much of this residual variation is explained by the introduction of another constant into the estimating equation. This is most easily done by: (1) subtracting 2.18371 (the variation explained by the second degree curve) from 3.85061 (Σy^2) in order to obtain the unexplained variation, 1 6690, after use of the second degree curve; (2) subtracting this amount (1.6690) from 1.82168 (the unexplained variation as measured from the straight line), giving .15478, the increment explained by the second degree curve.

Let us summarize these results also:

Source of variation	Amount of variation	Degrees of freedom	Variance
Increment explained by additional constant in equation	0 15478	1	0 15478
Unexplained by second degree curve	1.66690	14	11906
Total unexplained by straight line...........	1 82168	15	0.12145

A word of explanation concerning the determination of the degrees of freedom will be given. There are 17 items. However, since total variation is measured from the mean, one degree of freedom is lost. That is, arbitrary values may be assigned to any 16 of the 17 residuals, but the value of the other one is determined by the requirement that the deviations from the mean be zero. On the other hand, a straight line uses up an additional degree of freedom, or two altogether, since there are two constants in the equation (that is, a and b). Thus there are $17 - 2 = 15$ degrees of freedom remaining for the residuals from the straight line, as shown in the first table. This leaves $16 - 15 = 1$ degree of freedom for the y_c values. The second table shows 14 degrees of freedom for the deviations from the second degree curve, since the latter uses up three degrees of freedom on account of the constants a, b, and c. Deviations from their own mean of computed values obtained by the use of a second degree curve have 2 degrees of freedom, but as indicated in the above table, this is only 1 degree of freedom in addition to the single degree of freedom possessed by the deviations from their own mean of the computed values of a straight line.

Possibly the table on the next page will further clarify the nature of the different measures of variation into which the total variation has been divided.

Source of variation	Amount of variation	Degrees of freedom	Variance
Explained by straight line . ..	2.02893	1	2 02893
Increment due to additional constant	.15478	1	15478
Total explained by second degree curve	2.18371	2	1 09186
Unexplained by second degree curve	1 66690	14	.11906
Total..	3 85061	16	0 24066

We now want to ascertain if the explained variance attributable to the addition of a third constant is significant in relation to the remaining unexplained variance.   Thus

$$F = \frac{.15478}{.11906} = 1.300.$$

Our F table (Appendix G2) indicates that, when $n_1 = 1$ and $n_2 = 14$, the .05 level of significance requires that $F = 4.600$.   It is clear that the increase in correlation brought about by the addition of another constant to our equation is not significant.

Estimate of population correlation.   The best estimate of ρ for the population is

$$\bar{\rho}^2 = 1 - \frac{\bar{\sigma}^2_{y_s}}{\bar{\sigma}^2_y} = \frac{\rho^2(N - 1) - (m - 1)}{N - m}.$$

Using the first of the two expressions, since the variances needed are given above,

$$\bar{\rho}^2 = 1 - \frac{.11906}{.24066} = .5053.$$
$$\bar{\rho} = .711.$$

A similar correction for size of sample for r from the same data gives $\bar{r} = +.704$.   Although $\bar{\rho}$ is larger than $\bar{r}$, the difference is not great.   The result of this comparison should not be considered as conflicting with the analysis of variance test.   The mere failure to establish a significant difference does not prove that the difference is accidental.   It is still our best guess that the relationship between the two variables is curvilinear, but the linear hypothesis is by no means discredited.

Third degree curve.   As an illustration of the law of diminishing returns we shall use data derived from experiments with nitrogen fertilizer and

tobacco yield at Tipton, Georgia. One thousand pounds of fertilizer per acre were applied to five different plots. Of the active ingredients, phosphoric acid and potash were held constant at 8 per cent and 5 per cent respectively; and the nitrogen was made to vary as follows: none, 2 per cent, 3 per cent, 4 per cent, 5 per cent. Presumably the experiment was so conducted that differences in yield were not attributable to differences in soil fertility, drainage, etc., between plots. The experiment was repeated in three different years. Of the total variance, what proportion can be explained by the varying amount of nitrogen used? While it is possible that the experiment was not perfectly designed, the data indicate
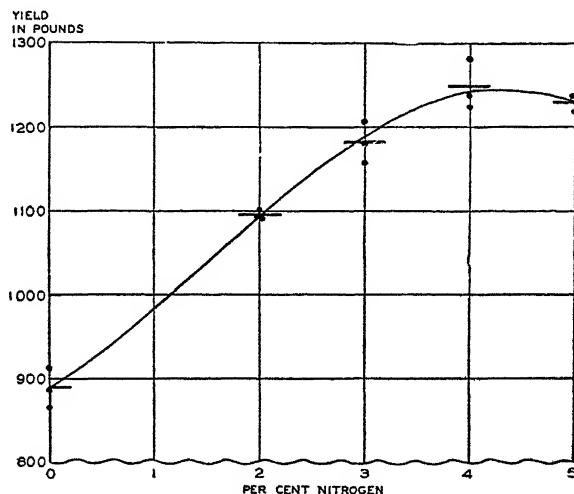
Chart 228. Per Cent Nitrogen in Fertilizer and Yield Per Acre of Tobacco, at Tipton, Georgia. (Horizontal lines indicate average yield per acre for each percentage of nitrogen, while curve represents computed values from equation $Y_c = 890.32389 + 78.263630X + 20.323899X^2 - 4.4648847X^3$. Data of Table 168.)

almost perfect correlation when the relationship is assumed to be of the equation type

$$Y_c = a + bX + cX^2 + dX^3.$$

This can be roughly verified by inspection of the scatter diagram, Chart 228. The heavy horizontal lines are the average yields for each of the percentages of nitrogen which are given. These means are not necessary for the solution of the problem, but are useful in discovering the type of curve to fit.

Solution of normal equations. Since four constants must be found, four normal equations of the following type must be used:[5]

$$\text{I.} \quad \Sigma Y = Na + b\Sigma X + c\Sigma X^2 + d\Sigma X^3;$$
$$\text{II.} \quad \Sigma XY = a\Sigma X + b\Sigma X^2 + c\Sigma X^3 + d\Sigma X^4;$$
$$\text{III.} \quad \Sigma X^2 Y = a\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4 + d\Sigma X^5;$$
$$\text{IV.} \quad \Sigma X^3 Y = a\Sigma X^3 + b\Sigma X^4 + c\Sigma X^5 + d\Sigma X^6.$$

The values required are computed in Table 168, and their substitutions result in the following normal equations:

$$\text{I.} \quad 16{,}934 = 15a + 42b + 162c + 672d;$$
$$\text{II.} \quad 50{,}630 = 42a + 162b + 672c + 2{,}934d;$$
$$\text{III.} \quad 197{,}198 = 162a + 672b + 2{,}934c + 13{,}272d;$$
$$\text{IV.} \quad 822{,}884 = 672a + 2{,}934b + 13{,}272c + 61{,}542d.$$

Following our previous procedure we may solve together equations I and II; II and III; III and IV, in each case eliminating a. We now have three equations:

$$\text{A.} \quad 48{,}222 = 666b + 3{,}276c + 15{,}786d;$$
$$\text{B.} \quad 80{,}256 = 1{,}980b + 14{,}364c + 82{,}116d;$$
$$\text{C.} \quad 790{,}152 = 23{,}724b + 178{,}416c + 1{,}051{,}020d.$$

We may now solve together A and B; B and C, eliminating b. The equations are thus reduced to two:

$$\text{D.} \quad -42{,}029{,}064 = 3{,}079{,}944c + 23{,}432{,}976d;$$
$$\text{E.} \quad -339{,}492{,}384 = 12{,}492{,}144c + 132{,}899{,}616d.$$

Solving equations D and E simultaneously, we find that $d = -4.4648847$ and $c = 20.323899$. By substituting these values in equation A, B, or C, we find that $b = 78.263630$. Substituting the values found for b, c, and d in equation I, II, III, or IV, we find a to have a value of 890.32389. It is advisable to check the values of d, c, b, and a at each step, since any error made in the early stages will vitiate all subsequent computations. One method of checking is to calculate each of the constants twice, by substituting in two different equations. Possibly even better is to substitute all of the constants known at any time in one of the remaining equations. For instance, if the value of a has been found by substituting

---

[5] Had observations been taken for 1 per cent nitrogen, the origin could conveniently have been taken at the mean of the X values (2.5). Then the sum of the odd powers of X would have been zero, and would have disappeared from the normal equations. We should then have had two pairs of normal equations to solve simultaneously:

$$\text{I.} \quad \Sigma Y = Na + c\Sigma X^2;$$
$$\text{III.} \quad \Sigma X^3 Y = a\Sigma X^2 + c\Sigma X^4.$$

$$\text{II.} \quad \Sigma XY = b\Sigma X^2 + d\Sigma X^4;$$
$$\text{IV.} \quad \Sigma X^2 Y = b\Sigma X^4 + d\Sigma X^6$$

The burden of computation would have been materially lightened.

TABLE 168

COMPUTATION OF VALUES REQUIRED TO OBTAIN MEASURES OF RELATIONSHIP BETWEEN PER CENT NITROGEN IN FERTILIZER AND YIELD PER ACRE OF TOBACCO, TIPTON, GEORGIA

(Fertilizer is 1,000 pounds per acre, P_2O_5 and K_2O are 8 per cent and 5 per cent respectively   The yields on all plots were unusually high in 1925; consequently they were reduced by a factor which reduced their average to the average of 1924 and 1926)

Plot number and year	Per cent nitrogen X	Yield in pounds Y	XY	X^2Y	X^3Y	X^2	X^3	X^4	X^5	X^6	Y^2
Plot 1:											
1924	0	867	0	0	0	0	0	0	0	0	751,689
1925	0	889	0	0	0	0	0	0	0	0	790,321
1926	0	914	0	0	0	0	0	0	0	0	835,396
Plot 2:											
1924	2	1,094	2,188	4,376	8,752	4	8	16	32	64	1,196,836
1925	2	1,101	2,202	4,404	8,808	4	8	16	32	64	1,212,201
1926	2	1,092	2,184	4,368	8,736	4	8	16	32	64	1,192,464
Plot 3:											
1924	3	1,206	3,618	10,854	32,562	9	27	81	243	729	1,454,436
1925	3	1,180	3,540	10,620	31,860	9	27	81	243	729	1,392,400
1926	3	1,157	3,471	10,413	31,239	9	27	81	243	729	1,338,649
Plot 4:											
1924	4	1,281	5,124	20,496	81,984	16	64	256	1,024	4,096	1,640,961
1925	4	1,238	4,952	19,808	79,232	16	64	256	1,024	4,096	1,532,644
1926	4	1,224	4,896	19,584	78,336	16	64	256	1,024	4,096	1,498,176
Plot 5:											
1924	5	1,235	6,175	30,875	154,375	25	125	625	3,125	15,625	1,525,225
1925	5	1,237	6,185	30,925	154,625	25	125	625	3,125	15,625	1,530,169
1926	5	1,219	6,095	30,475	152,375	25	125	625	3,125	15,625	1,485,961
Total	42	16,934	50,630	197,198	822,884	162	672	2,934	13,272	61,542	19,377,528

Source: W. J. Spillman, *Use of the Exponential Yield Curve in Fertilizer Experiments*, United States Department of Agriculture Technical Bulletin No. 348, pp. 16-17   Actually the five columns containing powers of X are not necessary   The quickest way to obtain these column totals is to look up the sums of the required powers of the first five natural numbers, subtract 1 (since X = 1 is missing), and multiply by 3 (since there are three years)

$$\bar{Y} = \frac{16,934}{15} = 1,128\ 933 \text{ pounds.}$$

values of b, c, and d in equation I, a final check may be made by substituting a, b, c, and d in equation IV. Thus

$$822,884 = 672(890.32389) + 2,934(78.263630) + 13,272(20.323899)$$
$$+ 61,542(-4.4648847)$$
$$= 598,297.65 + 229,625.49 + 269,738.79 - 274,777.93$$
$$= 822,884.00.$$

The estimating equation, then, is

$$Y_c = 890.32389 + 78.263630X + 20.323899X^2 - 4.4648847X^3.$$

Doolittle method. It must be confessed that, when there are as many as four equations to solve simultaneously, the above procedure is somewhat laborious Furthermore, no check can be applied until the value of d is obtained. Even that does not check the accuracy of any work except the solution of the two equations (D and E) necessary to obtain c and d. All of the preceding work could have been honeycombed with errors and still the solution of these two equations would check. It is not until all of the constants are obtained that we have any real check on the accuracy of the solution of the four normal equations. If the final check fails, all of the work must be repeated

Fortunately there is available for solving equations of this type simultaneously a systematic method that provides frequent checks on accuracy, and is less laborious than the above procedure when there are four or more equations It is known as the Doolittle method, having been developed by M H Doolittle. Like many labor-saving devices in statistics, the method at first seems very confusing. To a certain extent there is a substitution of complexity of procedure for repetitive drudgery.

The Doolittle method is illustrated by Table 169. There are five parts to this table:

Part 1. Normal equations. These are the same equations that are found on page 714, but all of the terms have been put on the left side, so that each equation equals zero.

Part 2. Forward solution. This solution obtains a value for d (-4 4648919, found in row IV', column Ω), and provides the figures with which to obtain values for the other constants.

Part 3. Back solution. In this part we compute by a simple process the values, in turn, for c, b, and a.

Part 4. Estimating equation Note that this equation agrees, to five digits, with the one previously obtained.

Part 5. Check equation. By substituting the values of the constants obtained in the last normal equation, the preceding work is checked. This step involves nothing new.

The entries in the forward solution are the most confusing, but if the procedure and explanation outlined below are followed very carefully, no trouble will be experienced in applying the Doolittle method to the solution of equations of this type. It is desirable that work be done in pencil first This will permit some of the entries to be made in boldface, as indicated in Table 169, merely by converting the pencil figures into ink The steps in the forward solution are as follows:

1. Divide the *forward solution* table into as many sections as there are normal equations. Leave a space between sections, and separate also by a horizontal line as shown Allow in each section two more rows than the section number: except that section one requires only two rows, rather than three.

2. Label the columns: (1), (2), (3), (4), Ω, and *Check total*   Five constants would require five normal equations, and therefore a column (5) also   Enter also the descriptive matter in the stub as shown in Table 169

3   Record the appropriate normal equation coefficients in the first row of each section, being sure to indicate minus signs.

4. Total each normal equation algebraically; record the results in the last column.

5. Make the following entries in the last row of each section:

> 1 00000000 in row   I′ column (1);
> 1.0000000 in row   II′ column (2);
> 1 000000   in row III′ column (3);
> 1 00000   in row IV′ column (4).

The number of zeros after 1 indicates the minimum number of decimal places to carry computations in each section.   The reason for dropping an additional decimal place as computations proceed from section to section is that errors from rounding the figures cumulate, and the number of significant places becomes smaller.   It is advisable, however, never to record fewer than eight digits, including the decimal places.

6. Row I′ is the result of dividing row I by the number in cell $\Sigma I(1)$ and changing signs.   The sum of the first five entries in this row should be checked against the entry in the total column, and agreement indicated by a check mark.   Values in columns (2), (3), (4), and Ω of this row should be entered in boldface, as further use is to be made of them.   (As suggested above, this is most easily done by reinforcing the original pencil entries with ink.)

7. The entries in the second row of section II, which is labeled $\Sigma I \times I'(2)$, are a result of multiplying the items in row ΣI by the number (in boldface) in the cell which is an intersection of row I′ and column (2).   In similar fashion, immediately below each row of normal equation coefficients are found the corresponding "product" rows These rows are called product rows because they are the result of making multiplications, a description of each such operation being given in the stub of the table.   It helps to keep the process straight if we observe that the multipliers are always the boldface numbers in the column bearing the same parenthesized number as the section being computed; and that the numbers multiplied are those in the row immediately above the boldface number in question   A check on the accuracy of these entries is afforded by totaling each row as it is computed, and indicating by a check mark agreement with the entry in the total column.

8. The third row of section II, labeled ΣII, is the result of adding algebraically the two rows above it in that section.   Likewise the Σ row in each section is a vertical summation of all the entries above the Σ row in the section in question   There is no separate Σ row in section I, since the section has no product row, and therefore the normal equation row automatically becomes also the Σ row.   Note that, as the computations proceed from section to section, there is an increase in the number of spaces in this row that are left vacant because the entries have become zero.   These Σ rows also should be added horizontally to obtain a check with the total column.

9. Row II′ is the result of dividing row ΣII by the value in $\Sigma II(2)$ and changing signs.   So also each "prime" row (III′, IV′, etc) is obtained by dividing each item in a given Σ row by the first entry in that row, with sign changed.   It is because of this fact that the first entry is always -1.   This entry is perhaps a sufficient description to remind us of the nature of the operation.   The prime rows should also check with the total column.   After the check has been made, enter the numbers to the right of each -1 in ink, up to, but not including, the total column entry.

TABLE 169

SOLUTION OF NORMAL EQUATIONS BY DOOLITTLE METHOD

(Normal equations from data of Table 168)

Part 1   Normal Equations

$$
\begin{aligned}
\text{I.} \quad & 15a + 42b + 162c + 672d - 16{,}934 = 0; \\
\text{II.} \quad & 42a + 162b + 672c + 2{,}934d - 50{,}630 = 0; \\
\text{III.} \quad & 162a + 672b + 2{,}934c + 13{,}272d - 197{,}198 = 0; \\
\text{IV.} \quad & 672a + 2{,}934b + 13{,}272c + 61{,}542d - 822{,}884 = 0.
\end{aligned}
$$

Part 2.   Forward Solution

Description of row	(1)	(2)	(3)	(4)	Ω	Check total
I and ΣI	15.	42	162	672.	− 16,934.	− 16,043.
I'.	− 1.00000000	− 2.80000000	− 10.80000000	− 44.80000000	− 1,128.93333333	− 1,069.53333333 √
II.	42.	162.	672	2,934.	− 50,630.	− 46,820.
ΣI × I'(2)	− 42.0000000	− 117.6000000	− 453.6000000	− 1,881.6000000	− 47,415.2000000	− 44,920.4000000 √
ΣII		44.4000000	218.4000000	1,052.4000000	3,214.8000000	1,899.6000000 √
II'.		− 1.0000000	− 4.9189189	− 23.7027027	− 72.4054054	− 42.7837838 √
III.	162.	672.	2,934.	13,272.	− 197,198.	− 180,158.
ΣI × I'(3)	−162.000000	− 453.600000	−1,749.600000	− 7,257.600000	182,887.200000	173,264.400000 √
ΣII × II'(3)		218.399999	−1,074.291888	−5,176.670250	15,813.340480	9,343.978342 √
ΣIII			110.108112	837.729750	1,502.540480	2,450.378342 √
III'.			1.000000	7.6082473	13.646047	22.254294 √
IV.	672.	2,934.	13,272.	61,542.	− 822,884.	− 744,464.
ΣI × I'(4)	−672.00000	−1,881.60000	−7,257.60000	−30,105.60000	758,643.20000	718,726.40000 √
ΣII × II'(4)		−1,052.40000	−5,176.67027	−24,944.72432	76,199.44864	45,025.65405 √
ΣIII × III'(4)			− 837.72975	−6,373.65511	−11,431.69955	−18,643.08440 √
ΣIV				118.02057	526.94909	644.96965 √
IV'.				1.00000	4.4648919	5.4648918 √

Part 3. Back Solution

Constant	Total (value of constant)	Boldface numbers in column immediately above multiplied by:				Computation of explained squares
		b	c	d	1	
a	890 32391	−219.137867	−219.498714	200.027157	1,128.93333333	15,076,745
b	78.263524	..	− 99.971886	105 830005	72 4054054	3,962,482
c	20.323955	...	...	33.970002	— 13 646047	4,007,843
d	— 4 4648919	...	...	..	— 4 4648919	−3,674,088
						$\Sigma Y_c^2 = 19{,}372{,}982$

Part 4. Estimating Equation

$$Y_c = 890\ 32391 + 78.263524X + 20\ 323955X^2 - 4\ 4648919X^3.$$

Part 5. Check (IV)

$$672(890.32391) + 2{,}934(78.263524) + 13{,}272(20.323955) + 61{,}542(-4\ 4648919) - 822{,}844 = .00.$$

The preceding explanation has referred specifically to the steps involved in sections I and II. The other sections are computed in similar fashion, each section requiring the previous computation of the other sections   The only variation among the different sections lies in the number of product rows and the number of vacant spaces to the left in some of the rows   As previously noted, we have obtained (in cell IV′Ω) the value of d which is -4.4648919 We are now ready to proceed with the back solution to obtain a, b and c.

The *back solution* occasions no difficulty   It consists merely in substituting the values of the constants, as obtained, in the derived equations III′, II′, and I′.   The entries in the 1 column are the boldface items in column Ω of the forward solution table.   The item in the last row of this column (-4.4648919) is d This value is recorded in the last row of the total column   The entries in the d column are the boldface items of column (4), above, multiplied by -4.4648919 (the value of d).   The sum of the items in the third row is c ($33.970002 - 13.646047 = 20.323955$), which is entered in the total column, opposite c.   The entries in the c column are the boldface items of column (3), above, multiplied by c.   The sum of the items in the second row is b The entry in the b column is the boldface entry in column (2), above, multiplied by b The sum of the items in the first row is a.   It will be noticed that, in using the back solution table, we record the column to the right first and then proceed to the left; and in the total column we proceed from bottom to top.   Proceeding in this fashion is rather unusual, but most convenient in this case.

The estimating equation arrived at by the Doolittle method,

$$Y_c = 890.32391 + 78.263524X + 20.323955X^2 - 4.4648919X^3,$$

agrees to at least five digits with the equation previously obtained on page 716.

In the right hand column of the Doolittle back solution table is provided a convenient place for computation of the explained sum of squares by the usual expression

$$\Sigma Y_c^2 = a\Sigma Y + b\Sigma XY + c\Sigma X^2Y + d\Sigma X^3Y.$$

Note also that ΣY, ΣXY, ΣX^2Y, and ΣX^3Y (with signs changed) are found in column Ω of the forward solution table, the first row of each section, in that order from top to bottom; while a, b, c, and d are arranged, in corresponding order, in the left-hand part of the back solution table.   The computations show that

$$\Sigma Y_c^2 = 19,372,982.$$

Using the equation previously obtained

$$Y_C = 890.32389 + 78.263630X + 20.323899X^2 - 4.4648847X^3,$$

the computation of Y_C values is as follows:

X	$a + bX$	cX^2	dX^3	Y_c (pounds)
0	890.324	0	0	890 32
1	968.588	20.324	— 4.465	984.45
2	1,046.851	81.296	— 35.719	1,092.43
3	1,125.115	182 915	—120.552	1,187.48
4	1,203.378	325.182	—285.753	1,242.81
5	1,281 642	508.097	—558.111	1,231.63

As can be seen from Chart 228, there is a point of inflection at about $1\frac{1}{2}$ per cent nitrogen, and the curve reaches a maximum of nearly 1,250 pounds shortly after the nitrogen reaches 4 per cent. These are, respectively, the points of diminishing marginal returns and diminishing total returns. How to locate these points more exactly is explained in Appendix B, section XXIII–2.

$$\Sigma Y_C^2 = a\Sigma Y + b\Sigma XY + c\Sigma X^2 Y + d\Sigma X^3 Y$$
$$= 890.32389(16,934) + 78.263630(50,630) + 20.323899(197,198)$$
$$- 4.4648847(822,884)$$
$$= 19,372,981.$$

Values of σ_{y_s} and ρ are obtained in the usual fashion:

$$\sigma_{y_s}^2 = \frac{\Sigma Y^2 - (a\Sigma Y + b\Sigma XY + c\Sigma X^2 Y + d\Sigma X^3 Y)}{N}$$

$$= \frac{\Sigma Y^2 - \Sigma Y_C^2}{N} = \frac{19,377,528 - 19,372,981}{15} = \frac{4,547}{15} = 303.1.$$

$$\sigma_{y_s} = 17.41 \text{ pounds.}$$

$$\rho^2 = \frac{\Sigma Y_C^2 - \bar{Y}\Sigma Y}{\Sigma Y^2 - \bar{Y}\Sigma Y} = \frac{19,372,981 - (1,128,933)(16,934)}{19,377,528 - (1,128,933)(16,934)}$$

$$= \frac{19,372,981 - 19,117,357}{19,377,528 - 19,117,357} = \frac{255,624}{260,171} = .9825.$$

$$\rho = .9912.$$

Grouped data. As an illustration of fitting a second degree curve to grouped data we shall take the relationship found to exist in East-Central Illinois between the yield per acre of broom corn and the man hours expended per ton in harvesting the crop. The data are shown in Table 170 and have been plotted in Chart 229. The horizontal line in each column is its mean. Inspection of the position of these lines reveals that labor costs decline rapidly at first as the quality of the land improves, but eventually tend to become constant. Although the use of reciprocals or logarithms might yield good results, for purposes of illustration we shall use a curve of the type

$$Y_C = a + bX + cX^2.$$

Examination of Chart 229 shows that b will be negative and c positive. To facilitate computation, we may designate the origin to be $X = 633.33$, $Y = 112.50$; and we shall compute the equation first in terms of class intervals. The X interval is 66.67, and the Y interval 25. With that origin and in those units, the estimating equation takes this form:

$$d'_{Y_C} = a + bd'_X + c(d'_X)^2,$$

TABLE 170

Correlation Table for Computation of Values Required for Measures of Relationship Between Tons per Acre and Man Hours per Ton Required in Harvesting Broom Corn

TONS OF BROOM CORN PER ACRE (X)

Class limits	Mid-value	133 34 to 199 99	200 00 to 266 66	266 67 to 333 33	333 34 to 399 99	400 00 to 466 66	466 67 to 533 33
		166 67	233 33	300 00	366 67	433 33	500 00
250 00 to 274 99	262 50	−42 294 1 −42 294	−36 216 1 −36 216				
225 00 to 249 99	237 50						
200.00 to 224 99	212 50						
175 00 to 199 99	187 50		−18 108 1 −18 108		−12 48 1 −12 48	−9 27 1 −9 27	
150 00 to 174 99	162 50			−10 50 1 −10 50		−6 18 1 −6 18	
125 00 to 149 99	137 50				−4 16 3 −12 48	−3 9 4 −12 36	−2 4 3 −6 12
100 00 to 124 99	112 50					0 0 1 0 0	0 0 7 0 0
75 00 to 99 99	87 50						2 −4 2 4 −8
50 00 to 74 99	62 50						
25 00 to 49 99	37 50						
d'_X		−7	−6	−5	−4	−3	−2
$(d'_X)^2$		49	36	25	16	9	4
f_X		1	2	1	4	7	12
$f_X d'_X$		−7	−12	−5	−16	−21	−24
$f_X(d'_X)^2$		49	72	25	64	63	48
$f_X(d'_X)^3$		−343	−432	−125	−256	−189	−96
$f_X(d'_X)^4$		2,401	2,592	625	1,024	567	192

MAN HOURS PER TON (Y)

Source: See Chart 229.

TABLE 170 (Continued)

533 34 to 599 99 / 566 67	600 00 to 666.66 / 633 33	666 67 to 733 33 / 700 00	733 34 to 799 99 / 766 67	800 00 to 866 66 / 833 33	866.67 to 933 33 / 900 00	d'_Y	f_Y	$f_Y d'_Y$	$f_Y(d'_Y)^2$
						6	2	12	72
						5	0	0	0
						4	0	0	0
						3	3	9	27
−2 2 / 1 / −2 2						2	3	6	12
−1 1 / 2 / −2 2	0 0 / 4 / 0 0	1 1 / 1 / 1 1				1	17	17	17
0 0 / 7 / 0 0	0 0 / 8 / 0 0	0 0 / 3 / 0 0	0 0 / 1 / 0 0	0 0 / 2 / 0 0	0 0 / 1 / 0 0	0	32	0	0
1 −1 / 4 / 4 −4	0 0 / 10 / 0 0	−1 −1 / 14 / −14 −14	−2 −4 / 3 / −6 −12	−3 −9 / 1 / −3 −9		−1	34	−34	34
		−2 −2 / 3 / −6 −6	−4 −8 / 3 / −12 −24	−6 −18 / 4 / −24 −72		−2	10	−20	40
		−3 −3 / 1 / −3 −3			−12 −18 / 1 / −12 −18	−3	2	−6	18
1	0	1	2	3	4	N 103		$\Sigma f_Y d'_Y$ −16	$\Sigma f_Y(d'_Y)^2$ 220
1	0	1	4	9	16				
14	22	24	7	7	2	N 103			
−14	0	24	14	21	8	$\Sigma f_X d'_X$ −32		$\Sigma f d'_X d'_Y$ −238	
14	0	24	28	63	32	$\Sigma f_X(d'_X)^2$ 482		$\Sigma f(d'_X)^2 d'_Y$ 662	
−14	0	24	36	189	128	$\Sigma f_X(d'_X)^3$ −1,058			
14	0	24	112	567	512	$\Sigma f_X(d'_X)^4$ 8 630			

723

where, as in earlier chapters, d' refers to a deviation from an arbitrary origin in terms of class intervals.

The correlation table shown as Table 170 is an extension of the form used in simple correlation (Table 161). It is slightly more complex on account of the additional values which must be computed in order to obtain the constants for a second degree curve. Specifically, the following additional values must be computed:

$$\Sigma f_X (d'_X)^3; \ \Sigma f_X (d'_X)^4; \ \Sigma f(d'_X)^2 d'_Y.$$

The numbers in the center of each cell indicate the number of farms that fall within the different cell boundaries. The upper left-hand corner of each cell represents the products of $d'_X d'_Y$, as in a simple correlation table. The first and third quadrants are positive, while the second and fourth are negative. The $(d'_X)^2$ row is useful in order to obtain the $(d'_X)^2 d'_Y$ products, which are recorded in the upper right-hand corner of each cell (note insert opposite). As the reader can easily verify, the first and second quadrants must be positive, and the third

$d'_X d'_Y$	$(d'_X)^2 d'_Y$
	f
$fd'_X d'_Y$	$f(d'_X)^2 d'_Y$

and fourth negative. The values in the lower left-hand and lower right-hand corners of each cell are the $fd'_X d'_Y$ and $f(d'_X)^2 d'_Y$ values respectively. They are obtained by multiplying the numbers in the upper corners by the cell frequency. Finally, in the extreme lower right-hand corner of Table 170 are recorded $\Sigma fd'_X d'_Y$, the sum of the $fd'_X d'_Y$ values; and also $\Sigma f(d'_X)^2 d'_Y$, the sum of the values which were recorded in the lower right-hand corner of each cell. In obtaining these totals, care must be exercised to add algebraically—that is, to add the plus values and subtract the minus values.

This general type of correlation table can be extended for use with curves of higher degree also, but the added complexity and additional space required ultimately limit the practicability of the device.

The normal equations for the estimating equation are:

I. $\quad \Sigma f_Y d'_Y = Na + b\Sigma f_X d'_X + c\Sigma f_X (d'_X)^2;$

II. $\quad \Sigma fd'_X d'_Y = a\Sigma f_X d'_X + b\Sigma f_X (d'_X)^2 + c\Sigma f_X (d'_X)^3;$

III. $\Sigma f(d'_X)^2 d'_Y = a\Sigma f_X (d'_X)^2 + b\Sigma f_X (d'_X)^3 + c\Sigma f_X (d'_X)^4.$

Making substitutions from Table 170, we have:

I. $\quad -16 = 103a - 32b + 482c;$

II. $\quad -238 = -32a + 482b - 1,058c;$

III. $\quad 662 = 482a - 1,058b + 8,630c.$

From these normal equations the estimating equation is

$$d'_{Y_c} = -.55290435 - .40268355d'_x + .058222561(d'_x)^2.$$

The origin of this equation is $\overline{X}_d$, $\overline{Y}_d$, and the units in which it is stated are X intervals and Y intervals—that is, the origin is $X = 633.33$, $Y =$
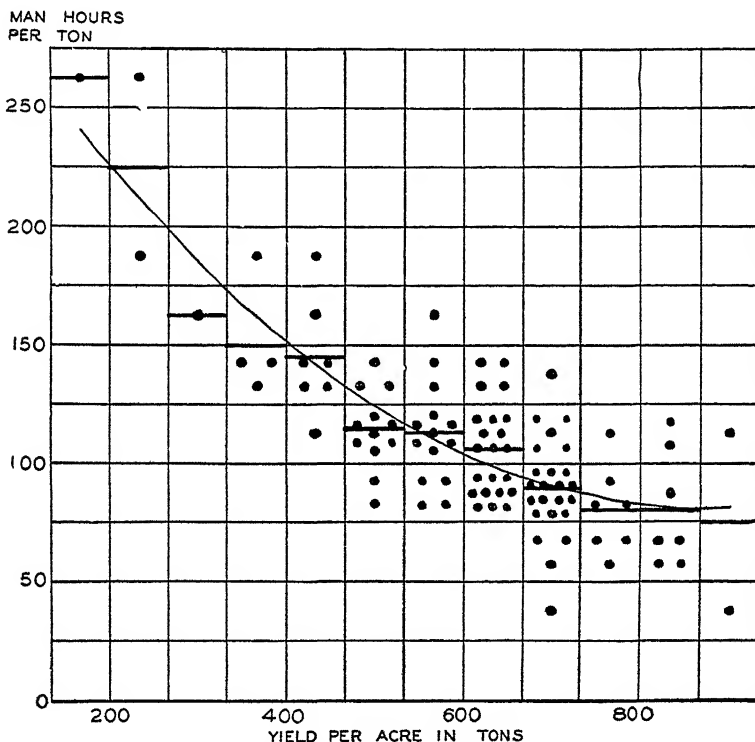
MAN HOURS
PER TON

Chart 229. Yield per Acre and Man-Hours per Ton Required to Harvest Broom Corn in East-Central Illinois. (Horizontal lines indicate average man-hours per ton for each yield, while curve represents computed values from equation $Y_c = 325.6794 - .5658420X + .0003275019X^2$. Data have been adapted from a chart on page 27 of *An Economic Study of Broom Corn Production*, by R. S. Washburn and J. H. Martin, U. S. Department of Agriculture, Technical Bulletin, No. 349, February 1933, p. 27.)

112.50, the X units 66.67 tons, and the Y units 25 man hours. Y_C values may be computed directly from this equation,[6] as shown by Table 171.

---

[6] If preferred, the formula can be transformed into the usual form. In this case it becomes

$$Y_C = 325.6794 - .5658420X + .0003275019X^2.$$

See Appendix B, section XXIII–3

TABLE 171

COMPUTATION OF Y_c VALUES FOR EQUATION

$$d'_{Y_c} = -.55290435 - .40268355\, d'_X + .058222561\, (d'_X)^2$$

X	d'_X	$(d'_X)^2$	$a + bd'_X$	$c(d'_X)^2$	d'_{Y_c} $[a + bd'_X + c(d'_X)^2]$	d_{Y_c} $[25d'_{Y_c}]$	Y_c $[\overline{Y}_d + d_{Y_c}]$ (man hours)
166 67	−7	49	2 265880	2 852905	5 118785	127 970	240.47
233 33	−6	36	1.863197	2 096012	3 959209	98 980	211 48
300 00	−5	25	1.460513	1 455564	2 916077	72 901	185 40
366 67	−4	16	1.057830	.931561	1 989391	49 735	162.24
433.33	−3	9	655146	524003	1 179149	29 479	141 98
500.00	−2	4	252463	232890	.485353	12 134	124 63
566 67	−1	1	− .150221	.058223	− 091998	− 2 300	110 20
633.33	0	0	− .552904	.000000	− .552904	−13 823	98 68
700 00	1	1	− .955588	058223	− .897365	−22 434	90 07
766.67	2	4	−1.358271	.232890	−1.125381	−28 135	84 37
833 33	3	9	−1.760955	.524003	−1 236952	−30 924	81 58
900 00	4	16	−2.163639	931561	−1 232078	−30 801	81 70

There is some doubt concerning the economic validity of this equation type for these data because of the fact that the curve begins to turn imperceptibly upward within the limits of the data.

To obtain σ_{y_s}, we use the expression

$$\sigma_{y_s} = i_Y \sqrt{\frac{\Sigma f_Y (d'_Y)^2 - \Sigma f_Y (d'_{Y_c})^2}{N}},$$

where $\Sigma f_Y (d'_{Y_c})^2 = a\Sigma f_Y\, d'_Y + b\Sigma f d'_X d'_Y + c\Sigma f (d'_X)^2 d'_Y$.

Substituting, we find

$$\Sigma f_Y (d'_{Y_c})^2 = (-.55290435)(-16) + (-.40268355)(-238)$$
$$+ (.058222561)(662)$$
$$= 143.22849;$$

and

$$\sigma_{y_s} = 25 \sqrt{\frac{220 - 143.22849}{103}} = 21.58.$$

ρ is perhaps most conveniently obtained as follows:

$$\rho^2 = \frac{\Sigma f_Y (d'_{Y_c})^2 - \dfrac{(\Sigma f_Y d'_Y)^2}{N}}{\Sigma f_Y (d'_Y)^2 - \dfrac{(\Sigma f_Y d'_Y)^2}{N}}$$

$$= \frac{143.22849 - \dfrac{(16)^2}{103}}{220 - \dfrac{(16)^2}{103}} = \frac{140.74305}{217.51456} = .6471.$$

$$\rho = .804.$$

Use of Means

Frequently it is difficult to decide, on theoretical or other grounds, the type of equation to choose. In such cases the statistician may select as a description of the relationship the average value of the dependent variable corresponding to different values of the independent variable. The computation of the measure of degree of relationship, usually called the *correlation ratio*, η, follows the same principle which we have hitherto used. It is the square root of the proportion of the total variation that has been explained by the variation of these average values. Two illustrations of procedure will be helpful.

A simple illustration. The nitrogen data provide an illustration that is simple in two respects. First, the $\overline{Y}_K$ values (that is, the column means) to be computed have been determined by the design of the experiment. We must compute the average yield when the per cent nitrogen is 0; 2,

3; 4; 5. Second, there are exactly three observations from which to compute each mean. The chief computations are shown in Table 172, which is divided into five main boxes (one for each value of X) and a total column.

The sums of the Y values corresponding to each percentage of fertilizer are recorded in row 1. Average yields are in row 2. These $\overline{Y}_K$ values are the explained values and, as may be seen from Chart 228, vary only slightly from the Y_C values previously computed by use of the equation $Y_C = a + bX + cX^2 + dX^3$. The explained sums of squares $\sum\limits_{1}^{m} N_K \overline{Y}_K^2$ is the sum of the squares of the individual column means multiplied by the number of observations in each column, or may be obtained more easily perhaps by the expression $\sum\limits_{1}^{m} \overline{Y}_K \sum\limits_{1}^{N_K} Y$. This value is computed in row 3, and found to be 19,373,216. The explained variation is obtained by subtracting the usual correction factor $\overline{Y}\Sigma Y$. ΣY is shown at the right of row 1 to be 16,934; hence $\overline{Y} = \dfrac{16,934}{15} = 1,128.9333$, and the correction factor is $1,128.93 \times 16,934 = 19,117,357$. The explained variation then is $19,373,216 - 19,117,357 = 255,859$. The total variation is computed exactly the same as with the other types of correlation we have considered. ΣY^2 is obtained from row 4. We may now find the measure of correlation:[7]

$$
\begin{aligned}
\eta^2 &= \frac{\sum\limits_{1}^{m}\left[N_K(\overline{Y}_K - \overline{Y})^2\right]}{\Sigma(Y - \overline{Y})^2} = \frac{\sum\limits_{1}^{m}\left(\overline{Y}_K \sum\limits_{1}^{N_K} Y\right) - \overline{Y}\Sigma Y}{\Sigma Y^2 - \overline{Y}\Sigma Y}, \\
&= \frac{19,373,216 - 19,117,357}{19,377,528 - 19,117,357} = \frac{255,865}{260,171}, \\
&= .9834. \\
\eta &= .9917.
\end{aligned}
$$

[7] It has been shown in Appendix B, section XIII–2, that

$$
\Sigma(Y - \overline{Y})^2 = \sum\limits_{1}^{m}\left[N_K(\overline{Y}_K - \overline{Y})^2\right] + \sum\limits_{1}^{m}\left[\sum\limits_{1}^{N_K}(Y - \overline{Y}_K)^2\right];
$$

$$
\Sigma Y^2 - \overline{Y}\Sigma Y = \left[\sum\limits_{1}^{m}\left(\overline{Y}_K \sum\limits_{1}^{N_K} Y\right) - \overline{Y}\Sigma Y\right] + \left[\Sigma Y^2 - \sum\limits_{1}^{m}\left(\overline{Y}_K \sum\limits_{1}^{N_K} Y\right)\right].
$$

That is to say, total variation equals variation between columns plus variation within columns; or, total variation equals explained variation plus unexplained variation. It is therefore obvious that η^2 may also easily be computed as 1 − (the ratio of the unexplained to the total variation).

In this text we are describing η_{YX} which uses the means of the columns in computing the explained variation. Sometimes the means of the rows are used instead, in which case the correlation ratio is denoted by η_{XY}. Although $r_{YX} = r_{XY}$, $\eta_{YX} \neq \eta_{XY}$.

TABLE 172

COMPUTATIONS REQUIRED TO COMPUTE CORRELATION RATIO BETWEEN PER CENT NITROGEN IN FERTILIZER AND YIELD PER ACRE OF TOBACCO

Year	X = 0		X = 2		X = 3		X = 4		X = 5		Total
	Y	Y²	Y	Y²	Y	Y²	Y	Y²	Y	Y²	
1924	867	751,689	1,094	1,196,836	1,206	1,454,436	1,281	1,640,961	1,235	1,525,225	⋮
1925	889	790,321	1,101	1,212,201	1,180	1,392,400	1,238	1,552,644	1,237	1,530,169	⋮
1926	914	835,396	1,092	1,192,464	1,157	1,338,649	1,224	1,498,176	1,219	1,485,961	⋮
Description of row:											
1. $\sum_1^{N_K} Y$	2,670		3,287		3,543		3,743		3,691		16,934
2. $\bar{Y}_K$	890 00		1,095 67		1,181 00		1,247 67		1,230 33		
3. $\bar{Y}_K \sum_1^{N_K} Y$	2,376,300 00		3,601,456.34		4,184,283 00		4,670,016 35		4,541,160 32		19,373,216
4. $\sum_1^{N_K} Y^2$		2,377,406		3,601,501		4,185,485		4,671,781		4,541,355	19,377,528

The similarity, as well as the difference, between the correlation ratio and the analysis of variance technique as treated in Chapter XIII should be mentioned. In the earlier chapter the ratio was between the explained and the unexplained population variance estimates, while here the ratio is between the explained and the total variation (or variance, if preferred) of the sample. Analysis of variance may be used to determine whether there is any significant relationship between the two variables; η^2 is used to measure the degree of relationship—the proportion of variation in the dependent variable, which has been explained by the independent variable.

Correlation in population. It should be recalled that for these data, $\rho = 9912$. This was increased to $\eta = .9917$ by the use of the five class means instead of the four constants involved in the estimating equation. Or, in terms of squared deviations, $\rho^2 = .9825$ and $\eta^2 = .9834$. This does not necessarily imply that the line of means more nearly represents the relationship which would be found in the population from which this sample was drawn, than does the four-constant estimating equation. Each time we add a constant to an equation, or subdivide the data so as to obtain another class mean, we reduce the possibility of variation from that estimating line or line of means. Each added constant or mean sacrifices a degree of freedom. In the present illustration, ρ sacrifices four degrees of freedom, while η sacrifices five. Now, the more complicated the relationship assumed, the less the variation which remains unexplained and the higher the apparent correlation. But since we have only a limited number of items in a sample, the results are misleading, for sacrificing a degree of freedom is equivalent to sacrificing an observation. When small samples are used, the scatter around the line of relationship is apt to be smaller than for the population, and therefore we shall tend to get correlation higher than exists in the parent population. However, we have seen that it is possible to make an estimate concerning the correlation which may reasonably be expected to obtain in the entire population, by allowing for the sample size and the degrees of freedom sacrificed. The easiest formula to use is

$$\bar{r}^2 = \frac{r^2(N-1) - (m-1)}{N-m},$$

where N is the number of items in the sample, and m is the number of constants in the equation or the number of class means—that is, the number of degrees of freedom sacrificed. This formula was discussed on page 679, and its derivation is shown in Appendix B, section XXII-6. It may be used not only for $\bar{r}$, but for $\bar{\rho}$, $\bar{\eta}$, and for the coefficient of multiple correlation discussed in the following chapter.[8]

---

[8] See Mordecai Ezekiel, *Methods of Correlation Analysis*, pp. 121-122, 246; John

Using this formula,

$$\bar{\rho}^2 = \frac{.9825(15 - 1) - (4 - 1)}{15 - 4} = \frac{.9825(14) - 3}{11} = .9777.$$

$$\bar{\rho} = .9888.$$

$$\bar{\eta}^2 = \frac{.9834(15 - 1) - (5 - 1)}{15 - 5} = \frac{.9834(14) - 4}{10} = .9768.$$

$$\bar{\eta} = .9883.$$

Thus we see that, by using class means instead of an equation, there has been no improvement in our explanation of the relationship between the per cent of nitrogen in the fertilizer and the yield of tobacco. Apparently there has been a slight retrogression!

Since the population estimate for ρ is greater than η, no purpose would be served by testing whether or not η is significantly greater than ρ. The procedure, however, is essentially the same as was used on pages 710–712 for testing the significance of the difference between ρ and r. We first discover the increase in the population variance brought about by using the line of means rather than the estimating equation. This explained variance is related to the unexplained variance by means of the z or F test. The unexplained variance is obtained from the deviations from column means. An example of this procedure will be given in connection with the next illustration.

It is sometimes desirable to test whether or not the given data exhibit a significant departure from linearity. The procedure is exactly the same as that described above. The increase in explained variance is that brought about by the use of the column means instead of the straight line equation.[9] In the present instance, however, there can be little doubt that the relationship between fertilizer and yield is non-linear.

The correlation ratio can be used not only when the independent variable is quantitative, but also when it is qualitative. Simon Kuznets has

---

Wiley and Sons, New York, 1930. The formula used above is the same as the one given by Ezekiel on p 121, but different symbols are used and the formula has been put in a form which probably is slightly easier to use.

[9] Another method of testing whether or not η is significantly greater than r is to compute $\eta^2 - r^2$ and compare this with its standard error.

$$\sigma_{\eta^2 - r^2} = 2\sqrt{\frac{\eta^2 - r^2}{N}} \sqrt{(1 - \eta^2)^2 - (1 - r^2)^2 + 1}.$$

The significance of this ratio may be roughly ascertained by referring to a table of normal curve areas. This test is not satisfactory, however, because it does not take into consideration the number of classes used in computing η nor does the sampling distribution of $\eta^2 - r^2$ always follow the normal curve.

suggested its use to test the validity of a seasonal index, which, it will be recalled, consists of means of columns of data, each column representing a separate month.[10]   This test is, of course, subject to the same limitations as is the analysis of variance test.   These limitations were mentioned on pages 497–498.

Data grouped on both axes. The broom corn data used earlier in this chapter provide an illustration that is more complex in three respects than the nitrogen data: (1) The number of classes and their limits are determined, not by the design of the experiment, but by the judgment of the statistician; (2) the number of items in the different classes vary; (3) the data have been grouped on the basis of man hours per ton as well as yield.

Since the data are grouped, we shall proceed to ascertain, first, the explained variation in intervals and, then, the total variation, also in intervals. η^2 will, of course, be the ratio of these two quantities. Table 173 is the computation table. As can be seen, there are thirteen main boxes in the body, one for each of the twelve classes, and one for the entire distribution. The box heading for each class indicates the class limits and mid-value of that class. The section for each of the twelve classes contains entries necessary to compute the class mean, while that for the entire series contains entries for computation of the standard deviation also. As the table shows, 112.5 man hours was arbitrarily taken as the origin of each column as well as for the entire Y series. Row 1 is for the totals of the various columns. Symbolically, these totals are N_K and $\overset{N_K}{\underset{1}{\Sigma}} f_Y d'_Y$ for columns corresponding to different X values; and N, $\Sigma f_Y d'_Y$, and $\Sigma f_Y (d'_Y)^2$ for the entire distribution. Rows 2 and 3 are necessary to obtain the explained sum of squares in intervals from the arbitrary origin. This value is 150.60065, and is recorded in the last column as the total of row 3.

The explained variation in terms of intervals is obtained by subtracting a correction factor from the explained sums of squared deviations as measured in deviations from the arbitrary origin. That is,

$$\overset{m}{\underset{1}{\Sigma}} \left(N_K \frac{\overline{Y}_K - \overline{Y}}{i} \right)^2 = \overset{m}{\underset{1}{\Sigma}} \left[\frac{\left(\overset{N_K}{\underset{1}{\Sigma}} f_Y d'_Y \right)^2}{N_K} \right] - \frac{(\Sigma f_Y d'_Y)^2}{N}$$

$$= 150.60065 - \frac{(16)^2}{103} = 148.11521.$$

---

[10] See Simon S. Kuznets, *Seasonal Variations in Industry and Trade*, p. 34n, National Bureau of Economic Research, New York, 1933.

TABLE 173

COMPUTATIONS REQUIRED TO COMPLETE CORRELATION RATIO BETWEEN YIELD PER ACRE OF BROOM CORN AND MAN HOURS PER TON REQUIRED FOR HARVESTING IN EAST-CENTRAL ILLINOIS

(Yield per acre in tons, X)

Cost in hours (Y)

Class	Mid-value	$d'y$	133.34–199.99 (160.67) fx	$fyd'y$	200.00–266.33 (233.33) fx	$fyd'y$	266.67–333.33 (300.00) fx	$fyd'y$	333.34–399.99 (366.67) fx	$fyd'y$	400.00–466.33 (433.33) fx	$fyd'y$	466.67–533.33 (500.00) fx	$fyd'y$	533.34–599.99 (566.67) fx	$fyd'y$	600.00–666.33 (633.33) fx	$fyd'y$	666.67–733.33 (700.00) fx	$fyd'y$	733.34–799.99 (766.67) fx	$fyd'y$	800.00–866.66 (833.33) fx	$fyd'y$	866.67–933.33 (900.00) fx	$fyd'y$	Entire distribution fx	$fyd'y$	$fy(d'y)^2$
250.00–274.99	262.50	5	1	5																							2	12	72
225.00–249.99	237.50	5																											
200.00–224.99	212.50	4			1	4	1																				3	9	27
175.00–199.99	187.50	3									3				1												3	6	12
150.00–174.99	162.50	2			1		2				2		1		2		4										17	17	17
125.00–149.99	137.50	1									1		7		2				1								17	17	17
100.00–124.99	112.50	–0									0		0		0		0		5	0	1	0	2	0	1	0	32	0	0
75.00–99.99	87.50	–1									1		2		4				14	–14	3	–3	1	–1			34	–34	34
50.00–74.99	62.50	–2											1						3	–6	3	–6	4	–8			10	–20	40
25.00–49.99	37.50	–3																	1	–3					1	–3	2	–6	18
Description of row.																													
1. Total			1	6	2	9	1	2	4	6	6	9	12		14	0	22		24	–22	7	–9	7	–9	2	–3	103	–16	220
2. $\frac{\sum_{1}^{N_K} fyd'y}{N_K}$			6.0000000		4.5000000		2.0000000		1.5000000		1.2857142		0.8333333		0.0000000		–2.7272730		–0.9166667		–1.2857142		–1.2857142		–1.5000000				
3. $\frac{\left(\sum_{1}^{N_K} fyd'y\right)^2}{N_K}$			36.00000		40.50000		4.00000		9.00000		11.57143		0.5333		0.0000		1.63636		20.16667		11.57143		11.57143		4.50000				

150 60085

The total variation in terms of intervals is[11]

$$\Sigma\left(\frac{Y - \bar{Y}}{i}\right)^2 = \Sigma f_Y (d'_Y)^2 - \frac{(\Sigma f_Y d'_Y)^2}{N}$$

$$= 220 - \frac{(16)^2}{103} = 217.51456.$$

For the ratio of determination, then, we have

$$\eta^2 = \frac{\displaystyle\sum_{1}^{m}\left[\frac{\left(\displaystyle\sum_{1}^{N_K} f_Y d'_Y\right)^2}{N_K}\right] - \frac{(\Sigma f_Y d'_Y)^2}{N}}{\Sigma f_Y (d'_Y)^2 - \frac{(\Sigma f_Y d'_Y)^2}{N}}$$

$$= \frac{150.60065 - 2.48544}{220.00000 - 2.48544} = \frac{148.11521}{217.51456} = .6809.$$

$$\eta = .825.$$

Comparison of ρ and η. The estimate for the entire population is

$$\bar{\eta}^2 = \frac{.6809(103 - 1) - (12 - 1)}{103 - 12} = .6423.$$

$$\bar{\eta} = .801.$$

Values of ρ^2 and ρ from these same data, using a second degree curve for the estimating equation, were .6471 and .804 respectively. Applying the same corrections, we find that

$$\bar{\rho}^2 = \frac{.6471(103 - 1) - (3 - 1)}{103 - 3} = .6400.$$

$$\bar{\rho} = .800.$$

It seems likely, therefore, that a second degree curve describes the true relationship between yield per acre of broom corn and man hours required for harvesting as accurately as does the line of means.

---

[11] These formulas are analogous to those used for ungrouped data and described on pages 354–355. It must be apparent that if we take

$$\Sigma\left(\frac{Y - \bar{Y}}{i}\right)^2 - \sum_{1}^{m}\left(N_K \frac{\bar{Y}_K - \bar{Y}}{i}\right)^2, \text{ we have}$$

$$\sum_{1}^{m}\left[\sum_{1}^{N_K}\left(\frac{Y - \bar{Y}_K}{i}\right)^2\right] = \Sigma f_Y (d'_Y)^2 - \sum_{1}^{m}\left[\frac{\left(\displaystyle\sum_{1}^{N_K} f_Y d'_Y\right)^2}{N_K}\right],$$

which is the unexplained variation in terms of intervals.

This impression is confirmed by an analysis of the variances obtained in computing ρ and η.   Recall that for these data:

$$\rho^2 = \frac{140.74305}{217.51456} = .6471; \text{ and } \rho = .804.$$

$$\eta^2 = \frac{148.11521}{217.51456} = .6809; \text{ and } \eta = .825.$$

In the table below are summarized the apportionment of the variation, degrees of freedom, and variance of this problem, each according to its source.

Source of variation	Amount of variation	Degrees of freedom	Variance
Explained by second degree curve .	140.74305	2	70.37152
Increment due to use of means	7.37216	9	.81913
Total explained by use of means.	148.11521	11	134.65019
Unexplained variation from means (variation within columns) . ..	69.39935	91	.76263
Total	217.51456	102	2.13250

The variation in row 2 (increment due to use of means) as well as the variation in row 4 (unexplained variation from means) is most easily obtained by subtractions within the table, although each can be obtained independently if we so desire.   In order to test the improvement in fit obtained by use of the line of means, we must relate the increment of explained variance due to the use of these means to the variance unexplained by these means.   Thus

$$F = \frac{.81913}{.76263} = 1.074.$$

Values of F are not stated in Appendix G2 for $n_1 = 9$ and $n_2 = 91$.   However, for the .05 level of significance, with $n_1 = 8$ and $n_2 = 60$, $F = 2.097$, which is greater than the computed F value.   Apparently considerably more than five times in one hundred random samples we should expect improvement as great·as that obtained.   Clearly the improvement is not significant.   Furthermore, a second degree curve is to be preferred to the line of means as an empirical description of the relationship between broom corn yield and man hours required for harvesting, since the relationship is simpler and indicates continuous, rather than discrete, changes in man hours per ton required to harvest varying yields per acre.

Limitations of correlation ratio.   The reader may already have been struck with certain rather obvious limitations to the usefulness of the cor-

relation ratio. In the first place, the data must be grouped according to some classification of the independent variable. In our nitrogen and crop yield illustration this grouping was determined by the design of the experiment, while in the broom corn illustration the grouping was somewhat arbitrary. In the second place, there is no estimating equation, but only the line of the means. Thus there is no hypothesis stated concerning the functional relationship between the variables, and no satisfactory way of making an estimate of the value of the dependent variable for any given value of the independent variable. Finally, the value of η approaches 1 as the number of columns is increased. This makes it especially important to estimate the value of η for the population. The formula for so doing, it will be remembered, takes into account not only the size of the sample, but the number of X groups into which the sample is divided.

Unreliability of Coefficients of Curvilinear Correlation

Approximate measures sometimes used to test the reliability of ρ and η are

$$\sigma_\rho = \frac{1 - \rho^2}{\sqrt{N - m}}.$$

$$\sigma_\eta = \frac{1 - \eta^2}{\sqrt{N - m}}.$$

In these formulae, m refers to the number of degrees of freedom sacrificed; that is, the number of constants in the fitted curve for ρ, or the number of X classes in the case of η. As in the case of r, the distribution of sample ρ's is approximately normal around ρ_P only if the sample is large and ρ_P is small. In addition to these restrictions the distribution of η is not normal unless the number of columns is indefinitely large. In fact, R. A. Fisher has pointed out that, for very large samples, $N\eta^2$ tends to be distributed as χ^2, with degrees of freedom equal to the number of columns minus 1.

A more rigorous test involves the analysis of variance. For ρ, we have

$$F = \frac{\bar{\sigma}_{y_c}^2}{\bar{\sigma}_{y_s}^2}.$$

In case (say) a third degree curve has been used as an estimating equation, answers to any of these questions may be found:

(1) Is the variance explained by a straight line significant?

(2) Is the additional variance explained by a second degree curve significant?

(3) Is the total variance explained by a second degree curve significant?

(4)   Is the additional variance explained by a third degree curve significant?

(5)   Is the total variance explained by a third degree curve significant?

In each case the unexplained variance (that considered as due to chance) is that remaining after use of the third degree curve.   As usual, n_1 is the degrees of freedom of the explained variance, and n_2 the degrees of freedom of the unexplained variance.   (If $\bar{\sigma}^2_{y_c} < \bar{\sigma}^2_{y_s}$, the explained variance is, of course, not significant.)

Notice that the unexplained variance is that remaining after making use of the estimating equation containing *all* the constants which have been computed, rather than that remaining after using the estimating equation which contains constants of no higher order than those being tested.   Thus, to answer question (3), the unexplained variation is that remaining after use of constant d rather than the unexplained variation remaining after use of constant c.   The latter quantity is the variation due, not to chance factors alone, but to chance factors plus constant d. If the latter quantity had been used to obtain the unexplained, or chance, variance, the F test might erroneously have failed to show significance for the constant being tested.   It may seem to the reader that use of the variance remaining after including d in the estimating equation, when it is the significance of c that is being tested, tends to force a showing of significance.   It is true that this procedure reduces the unexplained *variation*, but this fact is counteracted by the decrease in n_2, the number of degrees of freedom remaining.   This acts as an offsetting factor in two ways: (1) Since n_2 becomes smaller, the unexplained *variance* may actually become larger; (2) F must become larger for the same level of significance as n_2 decreases.

If an equation employing constant d has not been fitted, it is necessary to consider the variance unexplained by the second degree curve as the chance variance.   As mentioned before, there is some loss in accuracy in so doing, provided, of course, that the reduction of the unexplained variance by use of additional constants is not fortuitous.

When η has been computed, we may discover whether there is any significant correlation in the data by the use of

$$ F = \frac{\sum\limits_{1}^{m}\left[N_K(\bar{Y}_K - \bar{Y})^2 \right] \div n_1}{\sum\limits_{1}^{m}\left[\sum\limits_{1}^{N_K} (Y - \bar{Y}_K)^2 \right] \div n_2}. $$

If also first, second, and third degree curves have been fitted, each of the five questions above may be answered, as well as the additional question:

Is the additional variation explained by the line of means significant? To answer any of these questions, the explained variance (or additional explained variance) is compared with the variance remaining after use of the column means (that is, the variance within columns). To ask whether the additional variation explained by the line of means is significant is equivalent to asking whether the variation from the fitted curve has been significantly reduced, and therefore whether the fitted curve is a satisfactory hypothesis.

Selected References

Mordecai Ezekiel: *Methods of Correlation Analysis*, Chapters VI–IX inclusive; John Wiley and Sons, New York, 1930.

R. A. Fisher. *Statistical Methods for Research Workers* (Seventh Edition), pages 256–266; Oliver and Boyd, Edinburgh, 1938. Contains a discussion of $\sigma_{\eta^2 - r^2}$ and the fitness of the regression formula.

F. C. Mills: *Statistical Methods Applied to Economics and Business* (Revised Edition), Chapters XII, XV, XVII; Henry Holt and Co., New York, 1938. Chapter XII is on non-linear correlation. Chapter XV, pp. 501–521, illustrates the use of variance analysis in testing significance of correlation and of the equation type. Chapter XVII deals with the use of logarithms and reciprocals.

L. H. C. Tippett: *The Methods of Statistics* (Second Edition), Chapter IX; Williams and Norgate, Ltd., London, 1937. Includes tests of fitness of curve type by means of analysis of variance.

A. E. Waugh: *Elements of Statistical Methods*, Chapter X; McGraw-Hill Book Co., New York, 1938.

CHAPTER XXIV

MULTIPLE AND PARTIAL CORRELATION

Preliminary Explanation

Simple correlation. Before plunging into the theory of multiple and partial correlation it will be useful to review briefly the elementary principles of correlation, since the more refined measures involve simply an extension of the principles already discussed. First, an estimating equation of the type $Y_C = a + bX$ was computed by the method of least squares. This permitted us to make estimates of the value of the dependent variable from values of the independent variable. Next, it was demonstrated that the total variation of the dependent variable was the sum of the explained variation, and the variation which we had failed to explain by our hypothesis; that is, that $\Sigma y^2 = \Sigma y_C^2 + \Sigma y_S^2$. It should be remembered that we computed Σy^2 by the formula $\Sigma y^2 = \Sigma Y^2 - \bar{Y}\Sigma Y$; and that Σy_C^2 was computed by the expression $\Sigma y_C^2 = \Sigma Y_C^2 - \bar{Y}\Sigma Y$, in which $\Sigma Y_C^2 = a\Sigma Y + b\Sigma XY$ when we were dealing with simple correlation.

The standard error of estimate σ_{y_s}, which is $\sqrt{\dfrac{\Sigma y_S^2}{N}}$, enabled us to judge the range of error of our estimates of the dependent variable. Since

$$\Sigma y_S^2 = \Sigma y^2 - \Sigma y_C^2,$$
$$= \Sigma Y^2 - \Sigma Y_C^2,$$

σ_{y_s} can be calculated by a process which involves subtracting the explained variation from the total variation, or most easily from the expression

$$\sigma_{y_s} = \sqrt{\frac{\Sigma Y^2 - \Sigma Y_C^2}{N}}.$$

Finally, a measure was computed that permitted us to state the propor-

tion of total variation which had been explained by variations in the computed values of the dependent variable Y_C. This ratio

$$r^2 = \frac{\Sigma y_C^2}{\Sigma y^2} = \frac{\Sigma Y_C^2 - \overline{Y}\Sigma Y}{\Sigma Y^2 - \overline{Y}\Sigma Y}$$

was known as the coefficient of determination, and its square root was called the coefficient of correlation.

Multiple correlation. Exactly the same principles are involved in multiple correlation as in simple correlation, but the procedure is more laborious since there is more than one independent variable. Also, it is necessary to use slightly different symbols. The illustration in this chapter will deal with the relationship between suicide rates by regions, and average age, per cent male, and a business failure index in those same regions. Suicide rate is the dependent variable, and the other three are independent variables.

To simplify computations so that they can be shown in full in this chapter, the United States has been divided into 18 regions of substantially equal population and more or less homogeneous characteristics. With the exception of New York state, which has been divided into New York City and up-state New York, the boundaries of these regions follow state boundaries. The composition of the different regions can be observed by reference to Chart 230, which has been so drawn that equal areas on the map indicate equality of population. Selection of homogeneous areas of equal population serves to make the statistical results more reliable in that each region is given proper weight in the calculations, a consideration which statisticians frequently overlook in geographical correlations. On the other hand, use of only 18 observations with an equation of 4 constants does make the degrees of freedom dangerously small. The results obtained must therefore be regarded as primarily of illustrative importance.

It simplifies the notations somewhat if each of the variables is designated by the letter X, differentiating between the variables by means of subscripts. We shall designate our variables in this manner:

DEPENDENT VARIABLE:

Suicide rate .. .. X_1

INDEPENDENT VARIABLES:

Average age...... .... X_2
Per cent male   . .. . X_3
Business failure index. .... X_4

The first step in the correlation procedure is to obtain an equation which includes all three of these independent variables as a means of estimating a suicide rate for any region. The estimate is labeled $X_{C1.234}$, since it is

an estimate of variable X_1 computed from variables X_2, X_3, and X_4. Since there are three independent variables, there will be three b's. The equation type will be

$$X_{C1.234} = a_{1\ 234} + b_{12.34}\,X_2 + b_{13.24}\,X_3 + b_{14\ 23}\,X_4.$$

A word concerning the meaning of the b's and their subscripts is necessary. These *net coefficients of estimation* indicate the effect on X_1 of a change in the accompanying independent variable when allowance has been made for the other independent variables. Thus $b_{12.34}$ is an estimate of the variation in suicide rate associated with a variation in average age, independent of variation in per cent male or business failures. The social
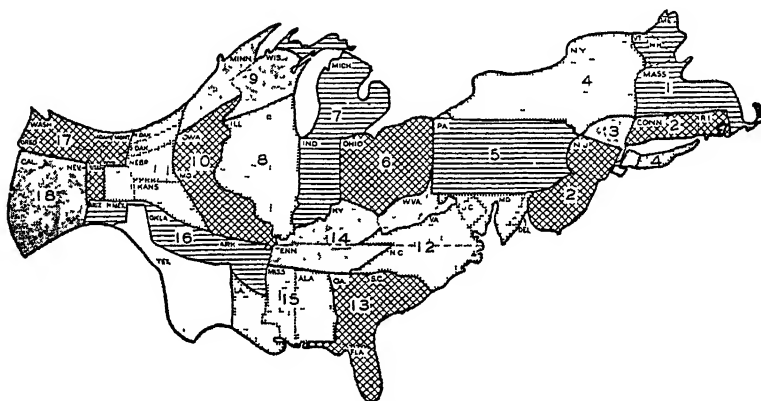
Chart 230.   Eighteen Regions of the United States of Substantially Equal Population and Homogeneous Characteristics.   (On this map the area of each state is proportional to its population.   Texas is not shaded, since it was not included in the death registration area in 1930.)

scientist is accustomed to saying "other things being equal." The other things which are held equal, *i.e.* at average value, are in this instance the proportion of males and the business failure rate in the different regions. As between regions that have the same per cent male and the same business failure rate, but differ with respect to age, each variation of one year in average age between regions will normally be accompanied by a variation of $b_{12.34}$ in suicide rate. The other b coefficients in the estimating equation are interpreted analogously, the figures to the right of the decimal point in the subscript indicating the factors that are held constant. Of course, really to know the effect on suicides of age alone, we should hold constant *all* other factors, not just per cent male and business failures. As we introduce more and more variables, this desirable situation is more and more closely approximated. The constant, $a_{1.234}$ is the hypothetical

value for suicide rate when the other factors considered have a value of zero. The estimate of, or normal value for, suicide rate of any region is the sum of the net amounts associated with each independent variable plus the value for a.

We might observe at this point that the natural scientist can often design his experiment so as to control a number of the variables, such for instance as temperature, humidity, or air pressure. The biologist and the agricultural experimenter can control their variables to a considerable extent. On the other hand, economics and sociology, and most of the social sciences, are observational, rather than experimental, sciences. Since workers in these fields usually have only a very limited control over the material they must use, they must attempt to hold some of the variables constant statistically, rather than experimentally—by means of the multiple correlation technique explained in this chapter.

As in previous instances the total variation of the dependent series is the sum of the variation in the estimated values of that series and the variation of the actual values from the estimated values, that is $\Sigma x_1^2 = \Sigma x_{C1\ 234}^2 + \Sigma x_{S1\ 234}^2$. The procedure in computing measures of relationship is essentially the same as with simple correlation. The standard error of estimate is

$$\sigma_{S1.234} = \sqrt{\frac{\Sigma x_{S1\ 234}^2}{N}},$$

and the *coefficient of multiple correlation is*

$$R_{1\ 234} = \sqrt{\frac{\Sigma x_{C1.234}^2}{\Sigma x_1^2}}.$$

$R_{1.234}^2$ states the proportion of total variation that is present in the variations of the computed, or $X_{C1\ 234}$, values, and which has been explained by reference to the independent variables. R has no sign, since the association may be positive with some, and negative with others, of the independent variables. It is interesting to note at this point that as additional associated independent variables are brought into the problem, $R_{1\ 234\ \cdots m}$ approaches 1.0 and $\sigma_{S1\ 234\ \cdots m}$ approaches zero, enabling us to make estimates which are progressively more accurate. If we were able to include all pertinent factors, we could make a perfect estimate and $R_{1\ 234\ \cdots m}$ would be 1.0.

Partial correlation. The coefficient of partial correlation (for example, $r_{14.23}$) is a measure of the relationship between the dependent variable and one independent variable, when the influence of the other independent variable (or variables) has theoretically been removed from both. The purpose of these coefficients is to show the relative importance of the different inde-

pendent variables in explaining variations in the dependent variable. This is done by finding the extent to which correlation is increased by the addition of another constant. More precisely, it may be said that the coefficient of partial determination (the square of the coefficient of partial correlation) is the ratio between *the increase in the variation of the computed values of the dependent variable resulting from introducing another independent variable (that is, the net variation associated with that factor)*, and *the variation that had not been explained before the introduction of the new factor*. The denominator of this ratio may also be regarded as the total variation which the new variable seeks to explain.

Turning now to our suicide illustration, a consideration of the two factors, average age X_2 and per cent male X_3, results in obtaining computed values for suicides, which we label $X_{C1\ 23}$. The variation that has been explained is indicated by the symbol $\Sigma x_{C1\ 23}^2$, and that which is still unexplained is $\Sigma x_{S1\ 23}^2$. If we now obtain an estimating equation by the use also of business failures as a third independent variable X_4, the explained variation becomes $\Sigma x_{C1\ 234}^2$. The increase in the explained variation is $\Sigma x_{C1\ 234}^2 - \Sigma x_{C1\ 23}^2$, and the coefficient of partial correlation is

$$r_{14\ 23} = \sqrt{\frac{\Sigma x_{C1\ 234}^2 - \Sigma x_{C1\ 23}^2}{\Sigma x_{S1\ 23}^2}}.$$

The subscript 14.23 indicates that the correlation is between suicide rate (X_1) and business failure rate (X_4) when average age (X_2) and per cent male (X_3) have been held constant. If we could pick out regions that are exactly alike with respect to age and per cent male, the simple correlation between suicide rate and business failure rate for those regions would tend to be the same as the above coefficient of partial correlation. The sign of $r_{14\ 23}$ is the same as that of $b_{14.23}$ in the estimating equation

$$X_{C1.234} = a_{1\ 234} + b_{12\ 34}\ X_2 + b_{13\ 24}\ X_3 + b_{14.23}\ X_4.$$

Computation Procedure

Computation of product sums. Since this chapter will require a considerable number of measures of relationship between the four variables, it will be convenient to compute at one time all the values that are needed in the different formulae. The computations of the sums and product sums are shown in Table 174.

There is one thing about this table which is worth special notice—the fact that there is an internal check on its accuracy. There are so many computations in multiple and partial correlation, each depending on the preceding, that it is foolhardy not to check each computation as it is done. To provide this check, a separate column is added to the first section of

TABLE 174

COMPUTATION OF PRODUCT SUMS REQUIRED FOR MEASURES OF RELATIONSHIP BETWEEN
SUICIDE RATES AND AVERAGE AGE, PER CENT MALE, AND BUSINESS FAILURES,
BY 18 REGIONS OF THE UNITED STATES, 1930

Region	Suicide rate X_1	Average age X_2	Per cent male X_3	Business failure rate X_4	Check column X_Σ
1	13.45	31 63	49 18	136.2	230.46
2	14 95	30 22	50 00	160 4	255.57
3	20.53	29 95	50 12	181 1	281.70
4	16 55	31 98	50 20	112 9	211 63
5	14 22	29 31	50.31	79 5	173.34
6	17.51	30.71	50 57	88.6	187 39
7	17.99	30.14	51 47	92 2	191.80
8	18 16	30 64	50 76	115 4	214 96
9	17 64	30.14	51.38	75 2	174.36
10	18.61	31.10	50 46	82 4	182.57
11	15 93	29 46	51.54	45 0	141 93
12	12 32	27 28	49 89	76 7	166 29
13	9 99	26 56	49 47	70 3	156 32
14	10 40	27 26	50 53	92 3	180 49
15	7 69	26 03	49 84	68 0	151 56
16	10 51	26.56	51 35	87.3	175 72
17	22 20	30 55	53 05	120 2	226 00
18	26 65	31.47	51.83	116 3	226 25
Total	285 30	531 09	911 95	1,800 0	3,528 34

Region	X_1^2	X_1X_2	X_1X_3	X_1X_4	Check column X_1X_Σ
1	180.9025	425 4235	661.4710	1,831.890	3,099.6870
2	223.5025	451 7890	747 5000	2,397.980	3,820.7715
3	421.4809	614.8735	1,028.9636	3,717.983	5,783.3010
4	273.9025	529 2690	830 8100	1,868.495	3,502.4765
5	202.2084	416.7882	715.4082	1,130.490	2,464 8948
6	306.6001	537.7321	885.4807	1,551.386	3,281.1989
7	323.6401	542.2186	925.9453	1,658.678	3,450.4820
8	329.7856	556.4224	921 8016	2,095.664	3,903.6736
9	311.1696	531.6696	906.3432	1,326.528	3,075.7104
10	346 3321	578.7710	939.0606	1,533.464	3,397.6277
11	253.7649	469.2978	821.0322	716.850	2,260 9449
12	151.7824	337.3216	614.6448	944.944	2,048.6928
13	99.8001	265.3344	494.2053	702.297	1,561.6368
14	108.1600	283.5040	525.5120	959.920	1,877.0960
15	59 1361	200.1707	383 2696	522.920	1,165.4964
16	110 4601	279.1456	539.6885	917.523	1,846.8172
17	492.8400	678.2100	1,177.7100	2,668.440	5,017.2000
18	710.2225	838.6755	1,381.2695	3,099.395	6,029.5625
Total	4,905.6904	8,536.6165	14,500 1161	29,644 847	57,587 2700

TABLE 174 (Continued)

COMPUTATION OF PRODUCT SUMS REQUIRED FOR MEASURES OF RELATIONSHIP BETWEEN
SUICIDE RATES AND AVERAGE AGE, PER CENT MALE, AND BUSINESS FAILURES,
BY 18 REGIONS OF THE UNITED STATES, 1930

Region	X_2^2	X_2X_3	X_2X_4	Check column X_2X_Σ
1	1,000.4569	1,555 5634	4,308 006	7,289.4498
2	913 2484	1,511 0000	4,847.288	7,723.3254
3	897 0025	1,501 0940	5,423.945	8,436.9150
4	1,022 7204	1,605 3960	3,610.542	6,767.9274
5	859 0761	1,474.5861	2,330 145	5,080.5954
6	943.1041	1,553 0047	2,720.906	5,754.7469
7	908 4196	1,551 3058	2,778 908	5,780 8520
8	938 8096	1,555 2864	3,535 856	6,586 3744
9	908 4196	1,548 5932	2,266 528	5,255.2104
10	967 2100	1,569 3060	2,562 640	5,677.9270
11	867 8916	1,518 3684	1,325 700	4,181 2578
12	749 6644	1,365 9882	2,100.046	4,553 0202
13	705 4336	1,313 9232	1,867.168	4,151 8592
14	743 1076	1,377 4478	2,516 098	4,920.1574
15	677 5609	1,297 3352	1,770 040	3,945.1068
16	705 4336	1,363 8560	2,318.688	4,667 1232
17	933 3025	1,620 6775	3,672.110	6,904.3000
18	990 3609	1.631 0901	3,659 961	7,120 0875
Total	15,731 2223	26,913.8220	53,614.575	104,796.2358

Region	X_3^2	X_3X_4	Check column X_3X_Σ
1	2,418 6724	6,698.316	11,334.0228
2	2,500 0000	8,020 000	12,778.5000
3	2,512 0144	9,076.732	14,118.8040
4	2,520 0400	5,667.580	10,623.8260
5	2,531.0961	3,999 645	8,720.7354
6	2,557.3249	4,480 502	9,476.3123
7	2,649.1609	4,745.534	9,871 9460
8	2,576 5776	5,857 704	10,911.3696
9	2,639 9044	3,863.776	8,958.6168
10	2,546.2116	4,157 904	9,212.4822
11	2,656.3716	2,319.300	7,315.0722
12	2,489 0121	3,826.563	8,296.2081
13	2,447.2809	3,477 741	7.733.1504
14	2,553.2809	4,663.919	9,120.1597
15	2,484 0256	3,389.120	7,553.7504
16	2,636.8225	4,482.855	9,023.2220
17	2,814.3025	6,376.610	11,989.3000
18	2,686.3489	6,027.829	11,726 5375
Total	46,218 4473	91,131.630	178,764.0154

TABLE 174 (Continued)

COMPUTATION OF PRODUCT SUMS REQUIRED FOR MEASURES OF RELATIONSHIP BETWEEN
SUICIDE RATES AND AVERAGE AGE, PER CENT MALE, AND BUSINESS FAILURES,
BY 18 REGIONS OF THE UNITED STATES, 1930

Region	X_4^2	Check column $X_4 X_\Sigma$
1	18,550 44	31,338 652
2	25,728 16	40,993.428
3	32,797 21	51,015 870
4	12,746 41	23,893 027
5	6,320 25	13,780 530
6	7,849 96	16,602 754
7	8,500 84	17,683 960
8	13,317 16	24,806 384
9	5,655.04	13,111 872
10	6,789 76	15,043 768
11	2,025 00	6,386 850
12	5,882 89	12,754.443
13	4,942 09	10,989.296
14	8,519 29	16,659.227
15	4,624 00	10,306 080
16	7,621 29	15,340 356
17	14,448 04	27,165.200
18	13,525.69	26,312,875
Total	199,843 52	374,234.572

Using the formulae given on page 747, the computations are checked as follows:

$$\Sigma X_\Sigma = 285.30 + 531\ 09 + 911.95 + 1,800.0$$
$$= 3,528\ 34$$
$$\Sigma X_1 X_\Sigma = 4,905\ 6904 + 8,536\ 6165 + 14.500.1161 + 29,644.847$$
$$= 57,587.2700.$$
$$\Sigma X_2 X_\Sigma = 8,536.6165 + 15,731.2223 + 26,913.8220 + 53,614.575$$
$$= 104,796.2358$$
$$\Sigma X_3 X_\Sigma = 14,500\ 1161 + 26,913\ 8220 + 46,218.4473 + 91,131.630$$
$$= 178,764.0154$$
$$\Sigma X_4 X_\Sigma = 29,644\ 847 + 53,614\ 575 + 91,131.630 + 199,843.52$$
$$= 374,234.572$$

Source: Computed from data found in publications listed below:

Average age United States Department of Commerce, Bureau of the Census, *Fifteenth Census of the United States, 1930*, Volume II

Per cent male United States Department of Commerce, Bureau of the Census, *Abstract of the Fifteenth Census of the United States, 1930*

Business failure rates United States Department of Commerce, Bureau of Foreign and Domestic Commerce, *Statistical Abstract of the United States, 1931*, and Dun and Bradstreet, Inc

Suicide rates United States Department of Commerce, Bureau of the Census, *Mortality Statistics, 1930* and *Abstract of the Fifteenth Census of the United States, 1930*.

this table, labeled X_Σ. Each item in this column, including its total, is the sum of the other items in the same row. If, therefore, the sum of the items in the X_Σ column equals the sums of the totals of columns X_1, X_2, X_3, X_4, the multiplications and additions are assumed to be correct. The right-hand column of each other section is also a check column. The checks are provided in each section by verifying the following identities:

$$\Sigma X_1 + \Sigma X_2 + \Sigma X_3 + \Sigma X_4 = \Sigma X_\Sigma.$$
$$\Sigma X_1^2 + \Sigma X_1 X_2 + \Sigma X_1 X_3 + \Sigma X_1 X_4 = \Sigma X_1 X_\Sigma.$$
$$\Sigma X_1 X_2 + \Sigma X_2^2 + \Sigma X_2 X_3 + \Sigma X_2 X_4 = \Sigma X_2 X_\Sigma.$$
$$\Sigma X_1 X_3 + \Sigma X_2 X_3 + \Sigma X_3^2 + \Sigma X_3 X_4 = \Sigma X_3 X_\Sigma.$$
$$\Sigma X_1 X_4 + \Sigma X_2 X_4 + \Sigma X_3 X_4 + \Sigma X_4^2 = \Sigma X_4 X_\Sigma.$$

By converting all of the product sums of Table 174 into deviations from the different means, the labor of computation will be materially lightened. This is because any straight line fitted by the method of least squares always passes through the means of the series and therefore a in the estimating equation becomes zero; and since there is one less constant to find, there is one less normal equation. To put the matter concretely, in our present problem, we may find directly the estimating equation

$$X_{C1\ 234} = a_{1\ 234} + b_{12\ 34} X_2 + b_{13\ 24} X_3 + b_{14.23} X_4,$$

which requires simultaneous solution of the four normal equations:

$$\Sigma X_1 = N a_{1\ 234} + b_{12\ 34}\Sigma X_2 + b_{13\ 24}\Sigma X_3 + b_{14\ 23}\Sigma X_4.$$
$$\Sigma X_1 X_2 = a_{1\ 234}\Sigma X_2 + b_{12\ 34}\Sigma X_2^2 + b_{13\ 24}\Sigma X_2 X_3 + b_{14\ 23}\Sigma X_2 X_4.$$
$$\Sigma X_1 X_3 = a_{1\ 234}\Sigma X_3 + b_{12.34}\Sigma X_2 X_3 + b_{13.24}\Sigma X_3^2 + b_{14.23}\Sigma X_3 X_4.$$
$$\Sigma X_1 X_4 = a_{1\ 234}\Sigma X_4 + b_{12\ 34}\Sigma X_2 X_4 + b_{13.24}\Sigma X_3 X_4 + b_{14\ 23}\Sigma X_4^2$$

A more expeditious procedure consists in using the estimating equation in terms of deviations. Thus

$$x_{C1.234} = b_{12\ 34} x_2 + b_{13.24} x_3 + b_{14.23} x_4.$$

Values for this equation are obtained by use of the normal equations in x_1, x_2, x_3, and x_4 instead of X_1, X_2, X_3, and X_4. Since $\Sigma x = 0$, the first normal equation disappears, the others becoming:

$$\Sigma x_1 x_2 = b_{12.34}\Sigma x_2^2 + b_{13.24}\Sigma x_2 x_3 + b_{14.23}\Sigma x_2 x_4.$$
$$\Sigma x_1 x_3 = b_{12\ 34}\Sigma x_2 x_3 + b_{13.24}\Sigma x_3^2 + b_{14.23}\Sigma x_3 x_4.$$
$$\Sigma x_1 x_4 = b_{12\ 34}\Sigma x_2 x_4 + b_{13.24}\Sigma x_3 x_4 + b_{14.23}\Sigma x_4^2.$$

To convert the product sums of Table 174 into deviation product sums we must subtract a correction factor from each:[1]

---

[1] The derivation of these equations is fairly obvious. The first and last will be taken as illustrations.
$$\Sigma x_1^2 = \Sigma(X_1 - \bar{X}_1)^2$$
$$= \Sigma(X_1^2 - 2\bar{X}_1 X_1 + \bar{X}_1^2)$$
$$= \Sigma X_1^2 - 2\bar{X}_1 \Sigma X_1 + N\bar{X}_1^2$$

$$\Sigma x_1^2 = \Sigma X_1^2 - \overline{X}_1 \Sigma X_1.$$
$$\Sigma x_2^2 = \Sigma X_2^2 - \overline{X}_2 \Sigma X_2;$$
$$\Sigma x_2 x_3 = \Sigma X_2 X_3 - \overline{X}_2 \Sigma X_3, \text{ or } \Sigma X_2 X_3 - \overline{X}_3 \Sigma X_2;$$
$$\Sigma x_2 x_4 = \Sigma X_2 X_4 - \overline{X}_2 \Sigma X_4, \text{ or } \Sigma X_2 X_4 - \overline{X}_4 \Sigma X_2;$$
$$\Sigma x_1 x_2 = \Sigma X_1 X_2 - \overline{X}_2 \Sigma X_1, \text{ or } \Sigma X_1 X_2 - \overline{X}_1 \Sigma X_2;$$
$$\Sigma x_3^2 = \Sigma X_3^2 - \overline{X}_3 \Sigma X_3;$$
$$\Sigma x_3 x_4 = \Sigma X_3 X_4 - \overline{X}_3 \Sigma X_4, \text{ or } \Sigma X_3 X_4 - \overline{X}_4 \Sigma X_3;$$
$$\Sigma x_1 x_3 = \Sigma X_1 X_3 - \overline{X}_3 \Sigma X_1, \text{ or } \Sigma X_1 X_3 - \overline{X}_1 \Sigma X_3;$$
$$\Sigma x_4^2 = \Sigma X_4^2 - \overline{X}_4 \Sigma X_4;$$
$$\Sigma x_1 x_4 = \Sigma X_1 X_4 - \overline{X}_4 \Sigma X_1, \text{ or } \Sigma X_1 X_4 - \overline{X}_1 \Sigma X_4;$$

These computations are made in Table 175, which has an internal check similar to that of Table 174. For instance, to verify Σx_4^2 and $\Sigma x_1 x_4$, we have:

$$\Sigma x_1 x_4 + \Sigma x_2 x_4 + \Sigma x_3 x_4 + \Sigma x_4^2 = \Sigma x_4 x_\Sigma;$$
$$505.575 - 63.370 + 19,843.52 + 1,114.847 = 21,400.572.$$

The diagram given on page 750 shows the method of making each check. In the diagram the dotted arrows indicate the product sums to be added in order to obtain the totals recorded in the x_Σ column.

Computation of gross measures of relationship. Simple correlation is in reality gross correlation, since it measures the relationship between two variables, without any adjustment by correlation technique for the effects of other variables. Using the symbols developed in the introductory section, we should compute the following measures if we wish to correlate suicide rates X_1 with average age X_2 alone:

Normal equations:

I. $\Sigma X_1 = N a_{1\,2} + b_{12} \Sigma X_2.$

II. $\Sigma X_1 X_2 = a_{1\,2} \Sigma X_2 + b_{12} \Sigma X_2^2,$ or $\Sigma x_1 x_2 = b_{12} \Sigma x_2^2.$

Estimating equation:

$X_{C1\,2} = a_{1\,2} + b_{12} X_2,$ or $x_{C1\,2} = b_{12} x_2.$

Sum of squares of computed values:

$\Sigma X_{C1\,2}^2 = a_{1.2} \Sigma X_1 + b_{12} \Sigma X_1 X_2,$ or $\Sigma x_{C1.2}^2 = b_{12} \Sigma x_1 x_2$
(Sum of explained squares)   (Explained variation)

$$= \Sigma X_1^2 - 2\overline{X}_1 \Sigma X_1 + \overline{X}_1 \Sigma X_1$$
$$= \Sigma X_1^2 - \overline{X}_1 \Sigma X_1$$
$$\Sigma x_1 x_4 = \Sigma[(X_4 - \overline{X}_4)(X_1 - \overline{X}_1)]$$
$$= \Sigma(X_1 X_4 - \overline{X}_4 X_1 - \overline{X}_1 X_4 + \overline{X}_1 \overline{X}_4)$$
$$= \Sigma X_1 X_4 - \overline{X}_4 \Sigma X_1 - \overline{X}_1 \Sigma X_4 + \overline{X}_1 N \overline{X}_4$$
$$= \Sigma X_1 X_4 - \overline{X}_4 \Sigma X_1 - \frac{\Sigma X_1 \Sigma X_4}{N} + \frac{\Sigma X_1 \Sigma X_4}{N}$$
$$= \Sigma X_1 X_4 - \overline{X}_4 \Sigma X_1.$$

TABLE 175

COMPUTATION OF DEVIATION PRODUCT SUMS REQUIRED FOR MEASURES OF RELATIONSHIP BETWEEN SUICIDE RATES AND AGE, PER CENT MALE, AND BUSINESS FAILURES, BY 18 REGIONS OF THE UNITED STATES, 1930

Sums and Means

$\Sigma X_1 = 285\ 3$	$\Sigma X_2 = 531\ 09$	$\Sigma X_3 = 911\ 95$	$\Sigma X_4 = 1{,}800\ 0$	$\Sigma X_\Sigma = 3{,}528\ 34$
$\bar{X}_1 = 15\ 85$	$\bar{X}_2 = 29\ 505$	$\bar{X}_3 = 50\ 6688889$	$\bar{X}_4 = 100\ 0$	$\bar{X}_\Sigma = 196\ 0188889$

Product Sums, Correction Factors, and Deviation Product Sums

$\Sigma X_1^2 = 4{,}905\ 6904$	$\Sigma X_1 X_2 = 8{,}536\ 6165$	$\Sigma X_1 X_3 = 14{,}500\ 1161$	$\Sigma X_1 X_4 = 29{,}644\ 847$	$\Sigma X_1 X_\Sigma = 57{,}587\ 2700$
$\bar{X}_1 \Sigma X_1 = 4{,}522\ 0050$	$\bar{X}_1 \Sigma X_2 = 8{,}417\ 7765$	$\bar{X}_1 \Sigma X_3 = 14{,}454\ 4075$	$\bar{X}_1 \Sigma X_4 = 28{,}530\ 000$	$\bar{X}_1 \Sigma X_\Sigma = 55{,}924\ 1890$
$\Sigma x_1^2 = 383\ 6854$	$\Sigma x_1 x_2 = 118\ 8400$	$\Sigma x_1 x_3 = 45\ 7086$	$\Sigma x_1 x_4 = 1{,}114\ 847$	$\Sigma v_1 x_2 = 1{,}663\ 0810$
	$\Sigma X_2^2 = 15{,}731\ 2223$	$\Sigma X_2 X_3 = 26{,}913\ 8220$	$\Sigma X_2 X_4 = 53{,}614\ 575$	$\Sigma X_2 X_\Sigma = 104{,}796\ 2358$
	$\bar{X}_2 \Sigma X_2 = 15{,}669\ 8104$	$\bar{X}_2 \Sigma X_3 = 26{,}907\ 0848$	$\bar{X}_2 \Sigma X_4 = 53{,}109\ 000$	$\bar{X}_2 \Sigma X_\Sigma = 104{,}103\ 6717$
	$\Sigma x_2^2 = 61\ 4119$	$\Sigma x_2 x_3 = 6\ 7372$	$\Sigma x_2 x_4 = 505\ 575$	$\Sigma r_2 r_2 = 692\ 5641$
		$\Sigma X_3^2 = 46{,}218\ 4473$	$\Sigma X_3 X_4 = 91{,}131\ 630$	$\Sigma X_3 X_\Sigma = 178{,}764\ 0154$
		$\bar{X}_3 \Sigma X_3 = 46{,}202\ 9335$	$\bar{X}_3 \Sigma X_4 = 91{,}195\ 000$	$\bar{X}_3 \Sigma X_\Sigma = 178{,}759\ 4258$
		$\Sigma x_3^2 = 15\ 5138$	$\Sigma x_3 x_4 = -63\ 370$	$\Sigma x_3 x_\Sigma = 4\ 5896$
			$\Sigma X_4^2 = 199{,}843.52$	$\Sigma X_4 X_\Sigma = 374{,}234\ 572$
			$\bar{X}_4 \Sigma X_4 = 180{,}000\ 00$	$\bar{X}_4 \Sigma X_\Sigma = 352{,}834\ 000$
			$\Sigma x_4^2 = 19{,}843\ 52$	$\Sigma x_4 x_\Sigma = 21{,}400\ 572$

x_1 x_2 x_3 x_4 x_2

$\Sigma x_1^2 = 383.6854$

$\Sigma x_1 x_2 = 118.8400$

$\Sigma x_2^2 = 61.4119$

$\Sigma x_1 x_3 = 45.7086$

$\Sigma x_2 x_3 = 6.7372$

$\Sigma x_3^2 = 15.5138$

$\Sigma x_1 x_4 = 1,114.847$

$\Sigma x_2 x_4 = 505.575$

$\Sigma x_3 x_4 = -63.370$

$\Sigma x_4^2 = 10,843.52$

$\Sigma x_1 x_2 = 1,633.0810$

$\Sigma x_2 x_2 = 692.5641$

$\Sigma x_3 x_2 = 4.5896$

$\Sigma x_4 x_2 = 21,400.572$

x_1 x_2 x_3 x_4

750
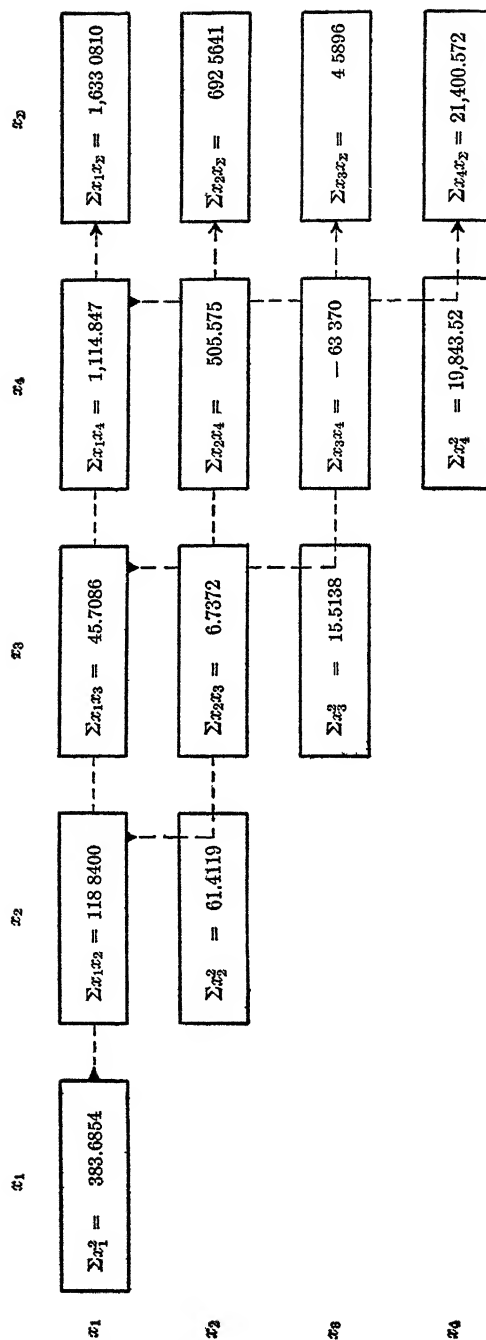
Sum of squares of deviations from estimates (unexplained variation):

$$\Sigma x_{S1\,2}^2 = \Sigma X_1^2 - \Sigma X_{C1\,2}^2, \qquad \text{or } \Sigma x_1^2 - \Sigma x_{C1\,2}^2$$

Standard error of estimate $\sqrt{\dfrac{\Sigma x_{S1\,2}^2}{N}}$:

$$\sigma_{S1\,2} = \sqrt{\frac{\Sigma X_1^2 - \Sigma X_{C1\,2}^2}{N}}, \qquad \text{or } \sqrt{\frac{\Sigma x_1^2 - \Sigma x_{C1.2}^2}{N}}.$$

Coefficient of correlation:

$$r_{12} = \sqrt{\frac{\Sigma X_{C1\,2}^2 - \overline{X}_1 \Sigma X_1}{\Sigma X_1^2 - \overline{X}_1 \Sigma X_1}}, \qquad \text{or } \sqrt{\frac{\Sigma x_{C1\,2}^2}{\Sigma x_1^2}}.$$

The careful reader will already have noticed that we are merely setting down the different equations and formulae used in simple correlation, with slightly different symbols. The coefficient of correlation r_{12} is sometimes called a zero order coefficient, since there are no additional independent variables held constant statistically.

Results of computations based on these expressions are given below; on the left, the data are taken in their original form, while on the right, deviations from means are used. All values are found in or derived from Table 175.

Normal equations:

I. $285.30 = 18a_{1\,2} + 531.09b_{12}.$

II. $8{,}536.6165 = 531.09a_{1.2} + 15{,}731.2223b_{12}.$ II. $118.840 = 61.4119b_{12}.$

Estimating equation:

$$X_{C1\,2} = -41.246051 + 1.9351314X_2. \qquad x_{C1\,2} = 1.93513x_2.$$

The equation $x_{C1.2} = 1.93513x_2$ may be converted into $X_{C1\,2} = -41.246 + 1.93513X_2$ by ascertaining the value of $a_{1.2}$ from the expression $a_{1\,2} = \overline{X}_1 - \overline{X}_2 b_{12}.$ Thus

$$a_{1.2} = 15.85 - (29.505)(1.93513) = -41.246.$$

Sums of explained squares: *Explained variation:*

$\Sigma X_{C1\,2}^2 = (-41.246051)(285.30)$ $\Sigma x_{C1\,2}^2 = (1.93513)(118.840)$

$\qquad\qquad + (1.9351314)(8{,}536.6165)$

$\qquad = 4{,}751.976.$ $= 229.971.$

Sum of squares of deviations from estimates:

$\Sigma x_{S1\,2}^2 = 4{,}905.6904 - 4{,}751.976$ $\Sigma x_{S1.2}^2 = 383.6854 - 229.971$

$\qquad = 153.7144.$ $= 153.7144.$

Standard error of estimate:

$$\sigma_{S1\,2}^2 = \frac{4{,}905.690 - 4{,}751.976}{18} \qquad \sigma_{S1.2}^2 = \frac{383.685 - 229.971}{18}$$

$\qquad = 8.540$ $= 8.540.$

$\sigma_{S1\,2} = 2.922$ suicides per 100,000. $\sigma_{S1\,2} = 2.922$ suicides per 100,000.
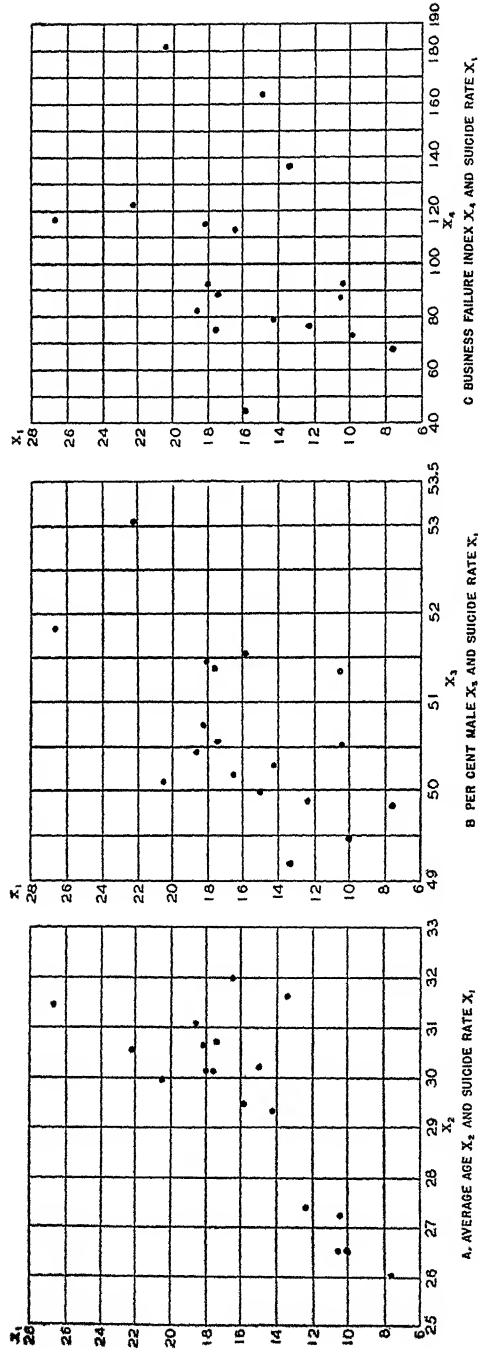
Chart 231. Scatter Diagrams of Gross Relationship of Average Age (X_2), Per Cent Male (X_3), and Business Failure Index (X_4), with Suicide Rate (X_1), in 18 Regions of the United States, 1930. (Data of Table 174.)

752

Coefficient of correlation:

$$r_{12}^2 = \frac{4,751.976 - (15.850)(285.30)}{4,905.690 - (15.850)(285.30)}. \qquad r_{12}^2 = \frac{229.971}{383.685}.$$

$$= \frac{4,751.976 - 4,522.005}{4,905.690 - 4,522.005} = .5994. \qquad = .5994.$$

$$r_{12} = +.7742. \qquad\qquad\qquad r_{12} = +.7742.$$

Chart 231 shows scatter diagrams of the simple relationship between suicide rates and each of the independent variables being considered. The standard errors of estimate and coefficients of simple correlation for these relationships are:

Suicide rate and average age:

$$\sigma_{S1.2} = 2.922 \text{ suicides per } 100,000; \qquad r_{12} = +.7742.$$

Suicide rate and per cent male:

$$\sigma_{S1\ 3} = 3.719 \text{ suicides per } 100,000; \qquad r_{13} = +.5925.$$

Suicide rate and business failure index:

$$\sigma_{S1\ 4} = 4.124 \text{ suicides per } 100,000; \qquad r_{14} = +.4040.$$

The evidence from these coefficients of correlation indicates that age is a fairly important factor bearing on suicide, and that per cent male and business affairs are of lesser importance in the order named. Age and per cent male are not necessarily *ultimate causes* of suicide, but the ultimate causes, whatever they are, seem to have a heavier incidence on men than on women, and on the old than on the young. On the other hand, recent studies have pointed to the conclusion that more women attempt suicide than men, but that men are more successful in killing themselves. Perhaps business failure may be thought of as a more fundamental cause of suicide. At any rate, economic factors are most commonly blamed by men who attempt suicide, while domestic difficulties are most commonly blamed by women.

Further information will be yielded by a careful study of Chart 232. Section A of this chart indicates the deviations of suicide rates from their mean, while section B shows the deviations in the estimates of suicide rate from their mean, that is, the individual explained variations. With several notable exceptions, the bars in this section appear about the same as in section A. Finally, section C indicates the individual variations that have not yet been accounted for; that is, the deviations of the actual suicide rates from the estimated rates. These deviations are obtained for each region by subtracting (algebraically) the value of the estimate from the actual value. Inspection of the chart will permit the reader roughly to verify the magnitude of the bars in section C. Since these unexplained
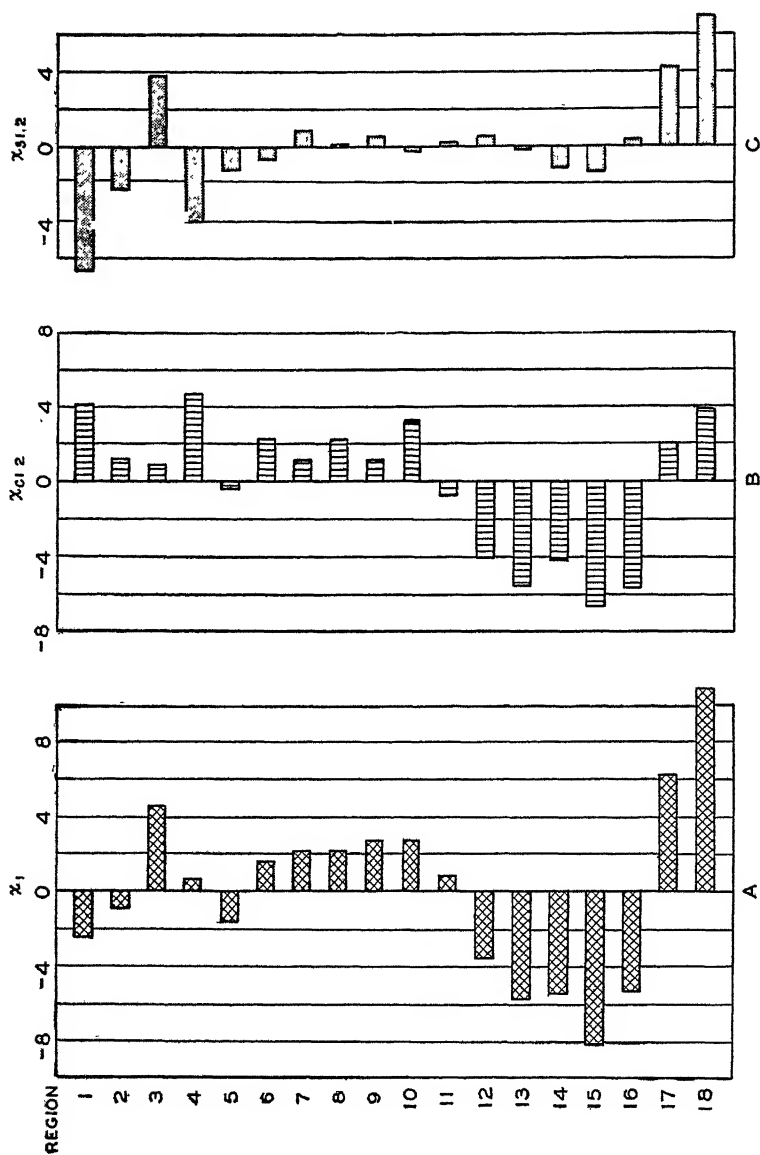
Chart 232. Deviations of Suicide Rates from Their Mean (x_1), Deviations from Their Mean of Computed Suicide Rates Based Upon Estimating Equation Using Average Age as the Independent Variable ($x_{C1.2}$), and Deviations of Suicide Rates from Computed rates ($x_{S1.2}$). (Derived from data of Table 174.)

variations are obtained by a subtraction process, they are often called *residuals*. The reader is already aware that, if the distances represented by each bar in this chart be squared, the sum of the squared values corresponding to sections B and C would equal those of section A.

In general, the bars in section C are much smaller than those in section A, but there are some exceptions. In the cases, for instance, of upper New England, North Atlantic, and up-state New York, it would have been more accurate to have guessed the suicide rates to have been 15.85, the simple average for the United States, than to have used the estimating equation. Confining ourselves now to the poorest estimates, we see from section C that we have yet to explain why the suicide rate was so low in

A. PER CENT MALE X_3 AND SUICIDE RATE
ADJUSTED FOR AVERAGE AGE $x_{S1\ 2}$

B. BUSINESS FAILURE INDEX X_4 AND SUICIDE RATE
ADJUSTED FOR AVERAGE AGE $x_{S1\ 2}$
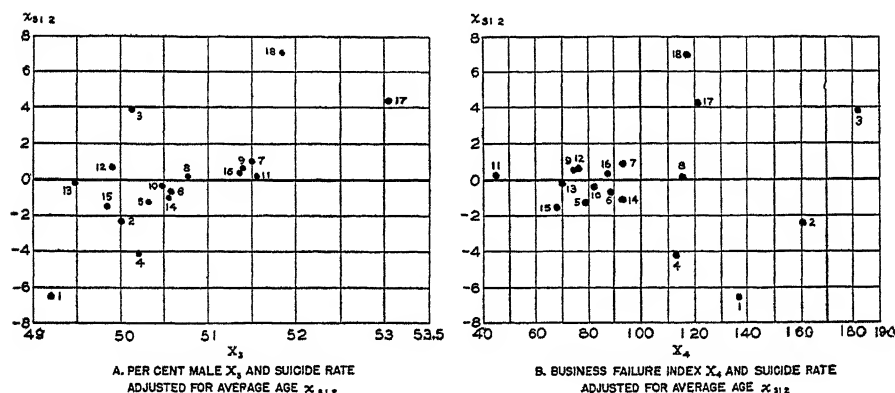
Chart 233. Scatter Diagrams of Per Cent Male (X_3) and Business Failure Index (X_4), Compared with Suicide Rate Adjusted for Average Age ($X_{S1\ 2}$). (Derived from data of Table 174.)

upper New England, North Atlantic, and up-state New York: and why so high in New York City, the Northwest states, and California.

Some clue to this difficulty is afforded by reference to Chart 233. In each section of this chart the dependent variable is the individual unexplained variations in suicide rate ($x_{S1\ 2} = x_1 - x_{C1.2}$) which were shown in section C of Chart 232. From section A of Chart 233 it is seen that regions 1, 2, and 4, which show large negative residuals in suicide rate, are low in per cent male also, while regions 17 and 18 are high, both with respect to positive residuals and per cent male. On the other hand, the per cent of males in New York City (region 3) appears to be below average. From section B of this chart we find that the business failure index number for New York City is exceptionally high, though business failures do not seem to explain very well the residuals in regions 1, 2, and 4. From an examination of the two sections of Chart 233 it is evident that we can

reduce the errors of our estimates and improve our correlation, more by including per cent male as a second factor than by including business failures. Consequently the next part of this chapter will correlate suicide rates with average age and per cent male simultaneously.

Before leaving this section, it is well to record values which were computed in connection with the simple correlation coefficients and which will be needed in subsequent sections.

Sum of squares:

These are not needed if all computations are made originally from deviations from means.

Total amount:

$$\Sigma X_1^2 = 4,905.690.$$

Amount explained by gross estimating equations:

$$\Sigma X_{C1.2}^2 = 4,751.976.$$
$$\Sigma X_{C1\ 3}^2 = 4,656.676.$$
$$\Sigma X_{C1\ 4}^2 = 4,584.639.$$

Sums of squared deviations (measures of variation):

These are obtained, by appropriate formulae, directly from the data in deviation form, or they may be obtained by subtracting the correction factor $\bar{X}_1 \Sigma X_1 = 4,522.005$ from each of the above expressions.

Total variation:

$$\Sigma x_1^2 = 383.685.$$

Variation explained by gross estimating equations:

$$\Sigma x_{C1.2}^2 = 229.971.$$
$$\Sigma x_{C1.3}^2 = 134.671.$$
$$\Sigma x_{C1\ 4}^2 = 62.634.$$

Two independent variables: multiple correlation. Naturally we can expect to estimate suicide rates more accurately if we take two independent variables into consideration, rather than only one. Hence let us make estimates from both average age and per cent male. The estimating equation type is

$$X_{C1.23} = a_{1\ 23} + b_{12.3}\,X_2 + b_{13\ 2}\,X_3$$

or, in terms of deviations,

$$x_{C1.23} = b_{12.3}\,x_2 + b_{13.2}\,x_3.$$

The 1.23 subscripts after X and a tell us that we are estimating values of X_1 (suicide rates) from variables X_2 (average age) and X_3 (per cent

male)   The first b tells the normal change in suicide rates associated with a unit change in average age for regions that have the same per cent male composition; the second b tells us the normal change in suicide rates associated with a unit change in per cent male for regions of the same average age.

The normal equations required are:

I. $\Sigma X_1 = Na_{1\ 23} + b_{12\ 3}\ \Sigma X_2 + b_{13\ 2}\ \Sigma X_3;$
II. $\Sigma X_1 X_2 = a_{1\ 23}\Sigma X_2 + b_{12\ 3}\Sigma X_2^2 + b_{13\ 2}\Sigma X_2 X_3;$
III. $\Sigma X_1 X_3 = a_{1\ 23}\Sigma X_3 + b_{12\ 3}\Sigma X_2 X_3 + b_{13\ 2}\Sigma X_3^2.$

Making the required substitutions, we have:

I. $285.30 = 18a_{1\ 23} + 531.09b_{12\ 3} + 911.95b_{13\ 2};$
II. $8{,}536.6165 = 531.09a_{1\ 23} + 15{,}731.5223b_{12\ 3} + 26{,}913.8220b_{13.2};$
III. $14{,}500.1161 = 911.95a_{1\ 23} + 26{,}913.8220b_{12\ 3} + 46{,}218.4473b_{13\ 2}.$

Solving these three equations gives

$$X_{C1\ 23} = -146.12082 + 1.6925398X_2 + 2.2112877X_3$$

Some labor may be saved if the normal equations are put in terms of deviations from the means.   In this case the first equation disappears, since Σx_1, Σx_2, and Σx_3 are each zero.   The equations are:

II. $\Sigma x_1 x_2 = b_{12\ 3}\Sigma x_2^2 + b_{13\ 2}\Sigma x_2 x_3:$
III. $\Sigma x_1 x_3 = b_{12\ 3}\Sigma x_2 x_3 + b_{13\ 2}\Sigma x_3^2.$

Solving these equations simultaneously:

II. $118.84 = 61.4119b_{12.3} + 6.7372b_{13.2};$
III. $45.7086 = 6.7372b_{12\ 3} + 15.5138b_{13.2}.$

We have

$$x_{C1.23} = 1.692539x_2 + 2.211297x_3.$$

These b values agree closely with those obtained before.   From the latter estimating equation $a_{1.23}$ is found by the expression[2]

$$a_{1\ 23} = \overline{X}_1 - b_{12\ 3}\overline{X}_2 - b_{13.2}\overline{X}_3$$
$$= 15.85 - (1.692539)(29.505) - (2.211297)(50.6639)$$
$$= -146.121.$$

The value for the explained sum of squares is obtained by an expression analogous to that derived in Appendix B, section XXII–1, equation 3:

$$\Sigma X_{C1.23}^2 = a_{1.23}\Sigma X_1 + b_{12\ 3}\Sigma X_1 X_2 + b_{13\ 2}\Sigma X_1 X_3$$
$$= (-146.12082)(285.30) + (1.6925398)(8{,}536.6165)$$
$$+ (2.2112877)(14{,}500.1161)$$
$$= 4{,}824.222.$$

---

[2] See Appendix B, section XXIV–1.

The explained variation may be obtained by subtracting $\bar{X}_1 \Sigma X_1 = 4{,}522.005$ from the above value; or, if the deviation form is preferred, the variation is computed directly:

$$\Sigma x_{C1\ 23}^2 = b_{12\ 3}\Sigma x_1 x_2 + b_{13.2}\Sigma x_1 x_3$$
$$= (1.692539)(118.840) + (2.211297)(45.7086)$$
$$= 302.217.$$

The measures of relationship are now computed in a fashion precisely similar to that employed when there was only one independent variable.

$$\sigma_{S1.23}^2 = \frac{\Sigma X_1^2 - \Sigma X_{C1.23}^2}{N} = \frac{4{,}905.690 - 4{,}824.222}{18} = \frac{81.468}{18} = 4.526.$$

$$\sigma_{S1\ 23} = 2.127.$$

$$R_{1\ 23}^2 = \frac{\Sigma X_{C1\ 23}^2 - \bar{X}_1 \Sigma X_1}{\Sigma X_1^2 - \bar{X}_1 \Sigma X_1} = \frac{4{,}824.222 - 4{,}522.005}{4{,}905.690 - 4{,}522.005} = \frac{302.217}{383.685} = .7877.$$

$$R_{1\ 23} = .8875.$$

When the data are in deviation form,

$$\sigma_{S1\ 23}^2 = \frac{\Sigma x_1^2 - \Sigma x_{C1\ 23}^2}{N} = \frac{383.685 - 302.217}{18} = \frac{81.468}{18} = 4.526.$$

$$\sigma_{S1\ 23} = 2.127.$$

$$R_{1.23}^2 = \frac{\Sigma x_{C1\ 23}^2}{\Sigma x_1^2} = \frac{302.217}{383.685} = .7877.$$

$$R_{1\ 23} = .8875.$$

This coefficient of multiple determination ($R_{1\ 23}^2$) is the proportion of total variation that is present in the computed, or $X_{C1.23}$, values, and which has therefore been explained by reference to variables X_2 and X_3; the coefficient of multiple correlation ($R_{1.23}$) is the square root of the proportion of variation in suicide rates between regions explained by reference to the values of average age and per cent male in the various regions.

In similar fashion we obtain the corresponding measures of relationship from other combinations of two of the independent variables. The three possible combinations are as indicated below.

Suicide rate with average age and per cent male:

$$X_{C1.23} = -146.12082 + 1.6925398 X_2 + 2.2112877 X_3;$$
$$x_{C1\ 23} = 1.69254 x_2 + 2.21129 x_3.$$
$$\Sigma X_{C1.23}^2 = 4{,}824.222;$$
$$\Sigma x_{C1.23}^2 = 302.217.$$
$$\sigma_{S1.23}^2 = 4.526;$$
$$\sigma_{S1.23} = 2.127.$$
$$R_{1.23}^2 = .7877;$$
$$R_{1.23} = .8875.$$

Suicide rate with average age and business failure index:

$$X_{C1\ 24} = -40.00222 + 1.863474X_2 + .00870415X_4;$$
$$x_{C1.24} = 1.86347x_2 + .00870415x_4.$$
$$\Sigma X_{C1\ 24}^2 = 4,753.134;$$
$$\Sigma x_{C1\ 24}^2 = 231.129.$$
$$\sigma_{S1\ 24}^2 = 8.474;$$
$$\sigma_{S1\ 24} = 2.911.$$
$$R_{1.24}^2 = .6025;$$
$$R_{1\ 24} = .7762.$$

Suicide rate with per cent male and business failure index:

$$X_{C1\ 34} = -153.82095 + 3.2177785X_3 + .06645785X_4;$$
$$x_{C1.34} = 3.21778x_3 + .0664578x_4.$$
$$\Sigma X_{C1\ 34}^2 = 4,743.184;$$
$$\Sigma x_{C1\ 34}^2 = 221.179.$$
$$\sigma_{S1.34}^2 = 9.028;$$
$$\sigma_{S1\ 34} = 3.005.$$
$$R_{1\ 34}^2 = .5765;$$
$$R_{1\ 34} = .7593.$$

It is to be noted that the two best combinations include the factor of average age; the two poorest, business failure index. This would suggest that age is the most important of the three factors having to do with suicide rates, and business failures the least important. Although this is the same rank in importance that was found when coefficients of simple (gross) correlation were used, such is not necessarily the case.

A visual impression of our progress is afforded by Chart 234, which shows: deviations of suicide rates from their mean (x_1); deviations from their mean of computed suicide rates, based upon the estimating equation using average age and per cent male as independent variables ($x_{C1.23}$); and deviations of suicide rates from computed rates ($x_{S1.23}$). The bars representing $x_{C1\ 23}$, which are in section B, are the individual explained variations, while those representing $x_{S1.23}$, which are in section C, are the individual unexplained variations. First it should be observed that the bars in section B of Chart 234 are somewhat longer than the corresponding bars in section B of Chart 232, and that they parallel more closely those of section A. In mathematical language, the explained variation has increased from $\Sigma x_{C1.2}^2 = 229.971$ and $\Sigma x_{C1.3}^2 = 134.671$ to $\Sigma x_{C1.23}^2 = 302.217$. Because this is true, the correlation increased from $r_{12} = +.7742$ and $r_{13} = +.5925$ to $R_{1.23} = .8875$. Likewise, of course, the unexplained variations represented by the bars in section C have been reduced somewhat. Correspondingly, the standard error of estimate has declined from $\sigma_{S1.2}$
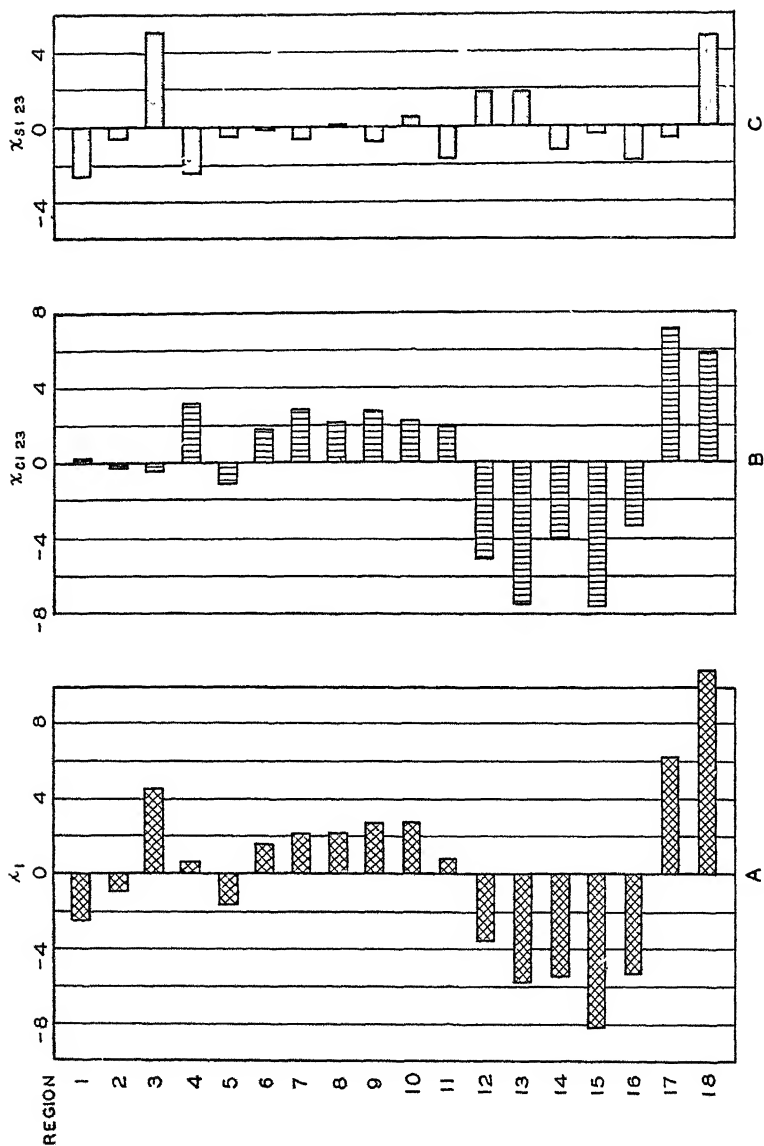
Chart 234. Deviations of Suicide Rates from Their Mean (x_1), Deviations from Their Mean of Computed Suicide Rates, Based upon Estimating Equation Using Average Age and Per Cent Male as Independent Variables ($x_{C1\ 23}$), and Deviations of Suicide Rates from Computed Rates ($x_{S1\ 23}$). (Derived from data of Table 174.)

= 2.912 and $\sigma_{S1\ 3} = 3.719$ to $\sigma_{S1\ 23} = 2.127$. It is always the case that, as more pertinent variables are introduced, the standard error of estimate becomes smaller, and the coefficient of multiple correlation larger. This is true even if some of the independent variables are negatively associated with the dependent variable. It is nevertheless true that suicide rates are considerably below our estimates for upper New England and up-state New York, and far above our estimates for New York City and California. In fact our estimate for New York City is worse than before. If we consult Chart 235, however, we shall see that the high suicide rates in these latter two regions (3 and 18) may partially be explained by business failures in those regions; but on the other hand, in upper New England (1) and up-state New York (4), where the lowness of the suicide rates is not accounted for, the business failure index is also higher than the United States average. It remains to be seen whether business failures, as such, are an important explanation of suicide. As judged by Chart 235, the relationship does not seem to be very close.
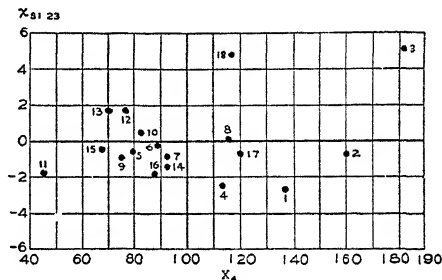
Chart 235. Scatter Diagram of Business Failure Index (X_4) and Suicide Rate Adjusted for Average Age and Per Cent Male ($x_{S1.23}$). (Derived from data of Table 174.)

Two independent variables: partial correlation. When only one independent variable (age) was considered, the deviations in our estimates were as shown in section B of Chart 232. By including an additional variable (per cent male) these explained deviations were increased to the amounts shown in Chart 234. In terms of symbols:

Variation explained by age and
per cent male . . . . $\Sigma x_{C1.23}^2 = \Sigma X_{C1.23}^2 - \overline{X}_1 \Sigma X_1$
Variation explained by age
alone . . . . $\Sigma x_{C1\ 2}^2 = \Sigma X_{C1.2}^2 - \overline{X}_1 \Sigma X_1$
Increase in variation explained ————————————————
by per cent male. . . . . $\Sigma x_{C1.23}^2 - \Sigma x_{C1.2}^2 = \Sigma X_{C1\ 23}^2 - \Sigma X_{C1.2}^2$

After taking age alone into consideration, the deviations remaining to be explained were those shown in Chart 232C. To summarize in terms of symbols:

$$\Sigma x_{S1.2}^2 = \Sigma X_1^2 - \Sigma X_{C1.2}^2, \text{ or } \Sigma x_1^2 - \Sigma x_{C1.2}^2.$$

The proportion of the variation previously unexplained, then, which was explained by including per cent male also, is the ratio

$$\frac{\Sigma X_{C1.23}^2 - \Sigma X_{C1\ 2}^2}{\Sigma X_1^2 - \Sigma X_{C1.2}^2},$$

or, if the deviation method is used,

$$\frac{\Sigma x_{C1.23}^2 - \Sigma x_{C1.2}^2}{\Sigma x_1^2 - \Sigma x_{C1\ 2}^2}.$$

This ratio is known as the *coefficient of partial determination*, the square root of which is the *coefficient of partial correlation*. Using the values already computed, therefore, we find:

$$r_{13.2}^2 = \frac{\Sigma X_{C1\ 23}^2 - \Sigma X_{C1\ 2}^2}{\Sigma X_1^2 - \Sigma X_{C1\ 2}^2} = \frac{4,824.222 - 4,751.976}{4,905.690 - 4,751.976} = \frac{72\ 246}{153.714} = .4700,\ \text{or}$$

$$= \frac{\Sigma x_{C1.23}^2 - \Sigma x_{C1\ 2}^2}{\Sigma x_1^2 - \Sigma x_{C1.2}^2} = \frac{302.217 - 229.971}{383.685 - 229.971} = \frac{72.246}{153.714} = .4700.$$

$$r_{13\ 2} = +.6856.$$

The sign of this coefficient of partial correlation is taken from the sign of $b_{13\ 2}$ in the estimating equation. This coefficient is a measure of the closeness of relationship between suicide rate and per cent male when age has been held constant statistically; it is the simple correlation coefficient which would be expected for regions of the same average age.

As a companion measure to $r_{13.2}$, we should obtain the partial coefficient $r_{12.3}$, which measures the relationship between suicide rate and age when per cent male has been held constant. This is done by finding the increase in the variation of the computed values by using age and per cent male in our estimating equation rather than using per cent male alone. Thus:

$$r_{12.3}^2 = \frac{\Sigma X_{C1\ 23}^2 - \Sigma X_{C1\ 3}^2}{\Sigma X_1^2 - \Sigma X_{C1.3}^2} = \frac{4,824.222 - 4,656.676}{4,905.690 - 4,656.676} = \frac{167.546}{249.014} = .6728,\ \text{or}$$

$$= \frac{\Sigma x_{C1.23}^2 - \Sigma x_{C1\ 3}^2}{\Sigma x_1^2 - \Sigma x_{C1\ 3}^2} = \frac{302.216 - 134.671}{383.685 - 134.671} = \frac{167.545}{249.014} = .6728.$$

$$r_{12\ 3} = +.8202.$$

The gross, or simple correlation between suicides and age, it will be recalled, was +.774. Removing the effect of variations in per cent male from both variables has increased the relationship materially—to +.820. Perhaps, however, the reader will be surprised to find a coefficient of multiple correlation of only .888, and coefficients of partial correlation

of $+.686$ and $+.820$. It is not a characteristic of these types of measures that the multiple coefficient is the sum of the two partial coefficients. The relationship is more complex than that.[3] It may be said, however, that for given values of r_{12} and r_{13} having the same sign the less the duplication in the independent variables (and so, the lower their positive, or the higher their negative, correlation; r_{23} in this case), the higher will be the multiple correlation.[4] In the present instance $r_{23} = +.218$, and hence the addition of either age or per cent male materially improves the estimate over that obtained from either alone. To aid the reader in seeing the interrelationships among the independent variables, scatter diagrams are shown in Chart 236, together with the correlation coefficients r_{23}, r_{24}, and r_{34}.

Other multiple and partial coefficients are:

$$R^2_{1\ 24} = \frac{\Sigma X^2_{c1.24} - \bar{X}_1 \Sigma X_1}{\Sigma X^2_1 - \bar{X}_1 \Sigma X_1} = \frac{4,753.164 - 4,522.005}{4,905.690 - 4,522.005} = \frac{231\ 159}{383.685} = .6024, \text{ or}$$

$$= \frac{\Sigma x^2_{c1\ 24}}{\Sigma x^2_1} = \frac{231.159}{383.685} = .6024.$$

$$R_{1\ 24} = .7762.$$

$$r^2_{14\ 2} = \frac{\Sigma X^2_{c1\ 24} - \Sigma X^2_{c1\ 2}}{\Sigma X^2_1 - \Sigma X^2_{c1\ 2}} = \frac{4,753.164 - 4,751.976}{4,905.690 - 4,751.976} = \frac{1.188}{153\ 714} = .007729, \text{ or}$$

$$\frac{\Sigma x^2_{c1\ 24} - \Sigma x^2_{c1\ 2}}{\Sigma x^2_1 - \Sigma x^2_{c1.2}} = \frac{231.159 - 229.971}{383\ 685 - 229.971} = \frac{1\ 188}{153.714} = .007729.$$

$$r_{14\ 2} = +.0879.$$

$$r^2_{12\ 4} = \frac{\Sigma X^2_{c1\ 24} - \Sigma X^2_{c1\ 4}}{\Sigma X^2_1 - \Sigma X^2_{c1\ 4}} = \frac{4,753.164 - 4,584\ 639}{4,905.690 - 4,584.639} = \frac{168.525}{321\ 051} = .5249, \text{ or}$$

$$\frac{\Sigma x^2_{c1\ 24} - \Sigma x^2_{c1\ 4}}{\Sigma x^2_1 - \Sigma x^2_{c1.4}} = \frac{231.159 - 62.634}{383.685 - 62\ 634} = \frac{168\ 525}{321.051} = .5249.$$

$$r_{12\ 4} = +.7245.$$

---

[3] The relationship is as follows:

$$R^2_{1.23} = \frac{r^2_{12} + r^2_{13} - 2r_{12}r_{13}r_{23}}{1 - r^2_{23}}.$$

In this case

$$R^2_{1.23} = \frac{.5994 + .3510 - 2(.7742)(.5925)(.2181)}{1 - .0476} = .7878.$$

$$R_{1.23} = .8876.$$

[4] However, if r_{12} and r_{13} have different signs, then, the lower the negative or the higher the positive correlation of r_{23}, the higher the value of $R_{1\ 23}$. The reader can verify these statements by assuming various values for r_{12}, r_{13}, and r_{23} and using the expression for $R_{1\ 23}$ given in footnote 3. Values of r_{23} inconsistent with the given values of r_{12} and r_{13} must not be used.
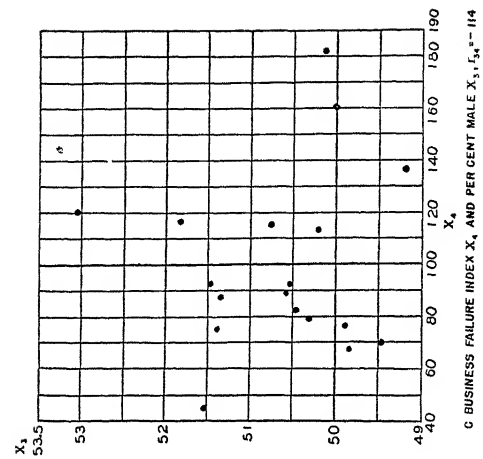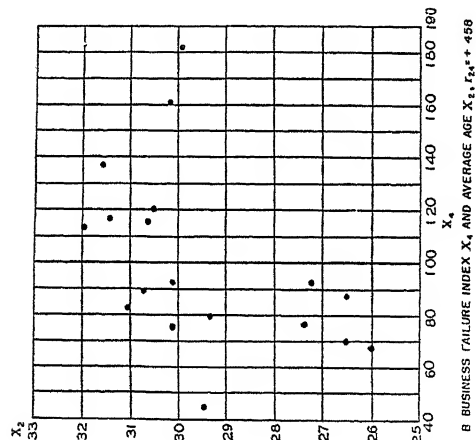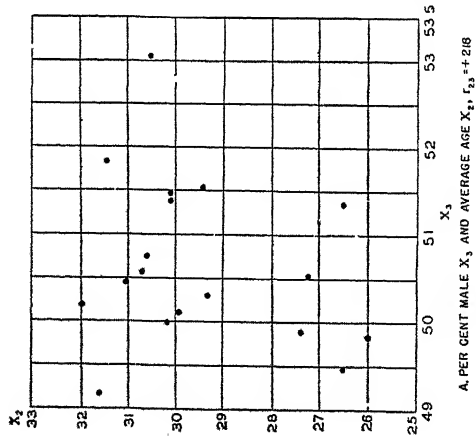
Chart 236. Simple or Gross Relationships among Average Age (X_2), Per Cent Male (X_3), and Business Failure Index (X_4), in 18 Regions of the United States, 1930. (Data of Table 174.)

$$R^2_{1\ 34} = \frac{\Sigma X^2_{c1\ 34} - \overline{X}_1\Sigma X_1}{\Sigma X^2_1 - \overline{X}_1\Sigma X_1} = \frac{4,743.184 - 4,522.005}{4,905.690 - 4,522.005} = \frac{221.179}{383.685} = .5765, \text{ or}$$

$$\frac{\Sigma x^2_{c1\ 34}}{\Sigma x^2_1} = \frac{221\ 179}{383.685} = .5765.$$

$$R_{1\ 34} = .7593$$

$$r^2_{14.3} = \frac{\Sigma X^2_{c1\ 34} - \Sigma X^2_{c1\ 3}}{\Sigma X^2_1 - \Sigma X^2_{c1.3}} = \frac{4,743.184 - 4,656.676}{4,905.690 - 4,656.676} = \frac{86.508}{249.041} = .3474, \text{ or}$$

$$\frac{\Sigma x^2_{c1\ 34} - \Sigma x^2_{c1\ 3}}{\Sigma x^2_1 - \Sigma x^2_{c1\ 3}} = \frac{221.179 - 134.671}{383.685 - 134.671} = \frac{86\ 508}{249.041} = 3474.$$

$$r_{14\ 3} = +.5894.$$

$$r^2_{13\ 4} = \frac{\Sigma X^2_{c1\ 34} - \Sigma X^2_{c1\ 4}}{\Sigma X^2_1 - \Sigma X^2_{c1.4}} = \frac{4,743.184 - 4,584.639}{4,905.690 - 4,584\ 639} = \frac{158.545}{321.051} = .4938, \text{ or}$$

$$\frac{\Sigma x^2_{c1\ 34} - \Sigma x^2_{c1\ 4}}{\Sigma x^2_1 - \Sigma x^2_{c1.4}} = \frac{221.179 - 62.634}{383\ 685 - 62.634} = \frac{158\ 545}{321.051} = .4938.$$

$$r_{13\ 4} = +.7027.$$

The results of computing these partial coefficients lead to the same conclusions as do the multiple coefficients. Looking at the partial r's, age is seen to be more closely related to suicides than is per cent male; per cent male more closely than business failures; and as might be supposed, age more closely than business failures.

It remains now to be seen whether the conclusions concerning the relative importance of our three independent variables will remain tenable when all four variables are considered simultaneously, rather than as different combinations of three variables. This problem will be considered in the following section. For the sake of simplicity, and since the reader should be sufficiently experienced with the longer procedure by now, data throughout the discussion will be used only in the form of deviations from means.

Three independent variables: multiple correlation. It is perhaps unnecessarily repetitive to go through with the same process again with one more variable added. The procedure is similar regardless of the number of variables. However, we have not yet definitely discovered how closely we can predict from all three independent variables, age, per cent male, and business failures; nor have we determined the relative importance of these factors.

The estimating equation and three normal equations required are as follows.

Estimating equation:

$$x_{c1.234} = b_{12.34}x_2 + b_{13.24}x_3 + b_{14\ 23}x_4.$$

Normal equations:

II. $\Sigma x_1 x_2 = b_{12\ 34}\Sigma x_2^2 + b_{13\ 24}\Sigma x_2 x_3 + b_{14\ 23}\Sigma x_2 x_4;$

III. $\Sigma x_1 x_3 = b_{12\ 34}\Sigma x_2 x_3 + b_{13\ 24}\Sigma x_3^2 + b_{14\ 23}\Sigma x_3 x_4;$

IV $\Sigma x_1 x_4 = b_{12\ 34}\Sigma x_2 x_4 + b_{13\ 24}\Sigma x_3 x_4 + b_{14\ 23}\Sigma x_4^2.$

If original data were used rather than deviations, four normal equations would be required (as shown on page 747); in such a case it would probably be advisable to use the Doolittle method of simultaneous solution, which was described on pp. 716–720. Inserting the appropriate values (found in Table 175) in the above three equations, we have:

II. $118.840 = 61.412 b_{12.34} + 6.737 b_{13\ 24} + 505.575 b_{14.23};$

III. $45.709 = 6.737 b_{12\ 34} + 15.514 b_{13\ 24} - 63.370 b_{14\ 23};$

IV $1{,}114.847 = 505.575 b_{12\ 34} - 63.370 b_{13\ 24} + 19{,}843.520 b_{14\ 23}.$

These equations solved simultaneously yield the estimating equation

$$x_{C1\ 234} = 1\ 445402 x_2 + 2.429389 x_3 + .02711406 x_4.$$

But

$$
\begin{aligned}
a_{1\ 234} &= \overline{X}_1 - b_{12\ 34}\overline{X}_2 - b_{13\ 24}\overline{X}_3 - b_{14\ 23}\overline{X}_4 \\
&= 15.85 - (1.445402)(29.505) - (2.429389)(50.66339) \\
&\qquad\qquad\qquad\qquad\qquad\qquad - (.02711406)(100) \\
&= -\ 152\ 589
\end{aligned}
$$

Therefore

$$X_{C1\ 234} = -152.589 + 1.445402 X_2 + 2.429389 X_3 + .02711406 X_4.$$

The variation of the computed values is

$$
\begin{aligned}
\Sigma x_{C1.234}^2 &= b_{12\ 34}\Sigma x_1 x_2 + b_{13.24}\Sigma x_1 x_3 + b_{14.23}\Sigma x_1 x_4 \\
&= (1.445402)(118.840) + (2.429389)(45.709) \\
&\qquad\qquad\qquad\qquad\qquad + (.02711406)(1{,}114.847) \\
&= 313{,}044.
\end{aligned}
$$

The other measures of relationship now are

$$\sigma_{S1\ 234}^2 = \frac{\Sigma x_1^2 - \Sigma x_{C1.234}^2}{N} = \frac{383{,}685 - 313.044}{18} = \frac{70.641}{18} = 3.924.$$

$$\sigma_{S1\ 234} = 1.981.$$

$$R_{1.234}^2 = \frac{\Sigma x_{C1.234}^2}{\Sigma x_1^2} = \frac{313.044}{383.685} = .8159.$$

$$R_{1.234} = .9033.$$

The coefficient of correlation has become progressively larger as the number of variables has been increased.   Thus

$$R_{1\,234} = .9033$$

$$R_{1.23} = .8875 \qquad R_{1.34} = .7593 \qquad R_{1.24} = .7761$$

$$r_{12} = +\,7742 \qquad r_{13} = +\,5925 \qquad r_{14} = +.4040 \qquad r_{12} = +.7742$$
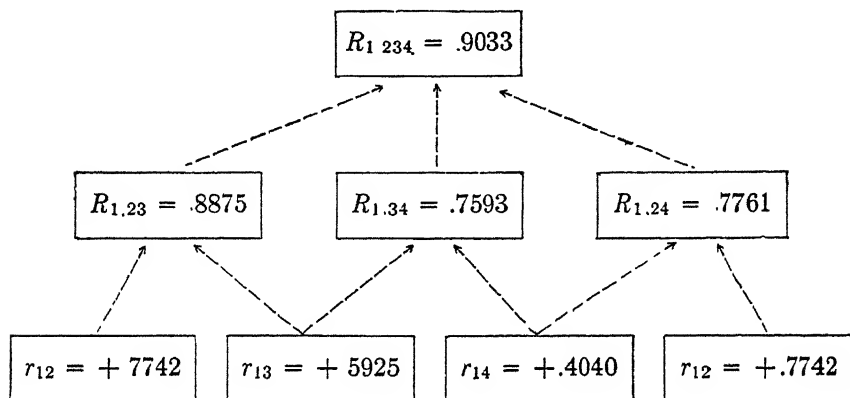
As explained earlier, neither the coefficients of correlation nor the coefficients of determination are additive to produce the higher multiple measures, on account of the duplication of elements involved.   The coefficients of correlation become larger and the standard errors of estimate become smaller as more factors are added, because the explained variance becomes larger and the unexplained variance becomes smaller.   The square root of the unexplained variance is, of course, σ_S.   The gradual reduction of σ_S as more factors are introduced is shown below.

$$\sigma_{S1.234} = 1.981$$

$$\sigma_{S1.23} = 2.127 \qquad \sigma_{S1.34} = 3.005 \qquad \sigma_{S1\,24} = 2.911$$

$$\sigma_{S1.2} = 2.922 \qquad \sigma_{S1.3} = 3\,719 \qquad \sigma_{S1.4} = 4.124 \qquad \sigma_{S1.2} = 2.922$$
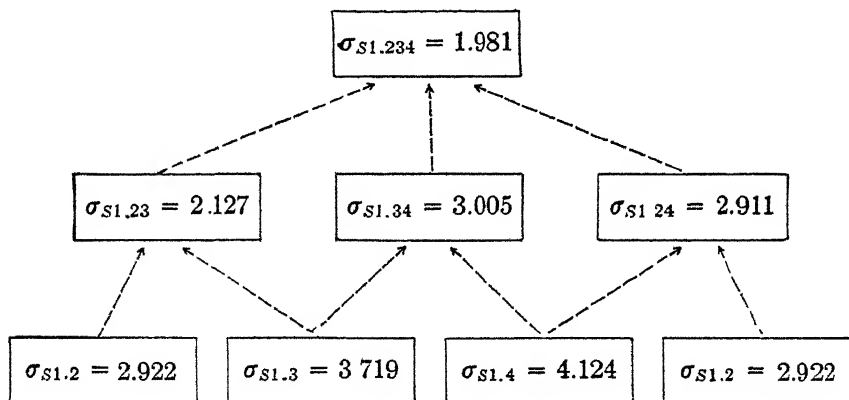
Chart 237, of the now familiar type, shows the deviations in suicide rates explained by our three independent variables and shows also the remaining unexplained deviations.   The addition of the economic factor
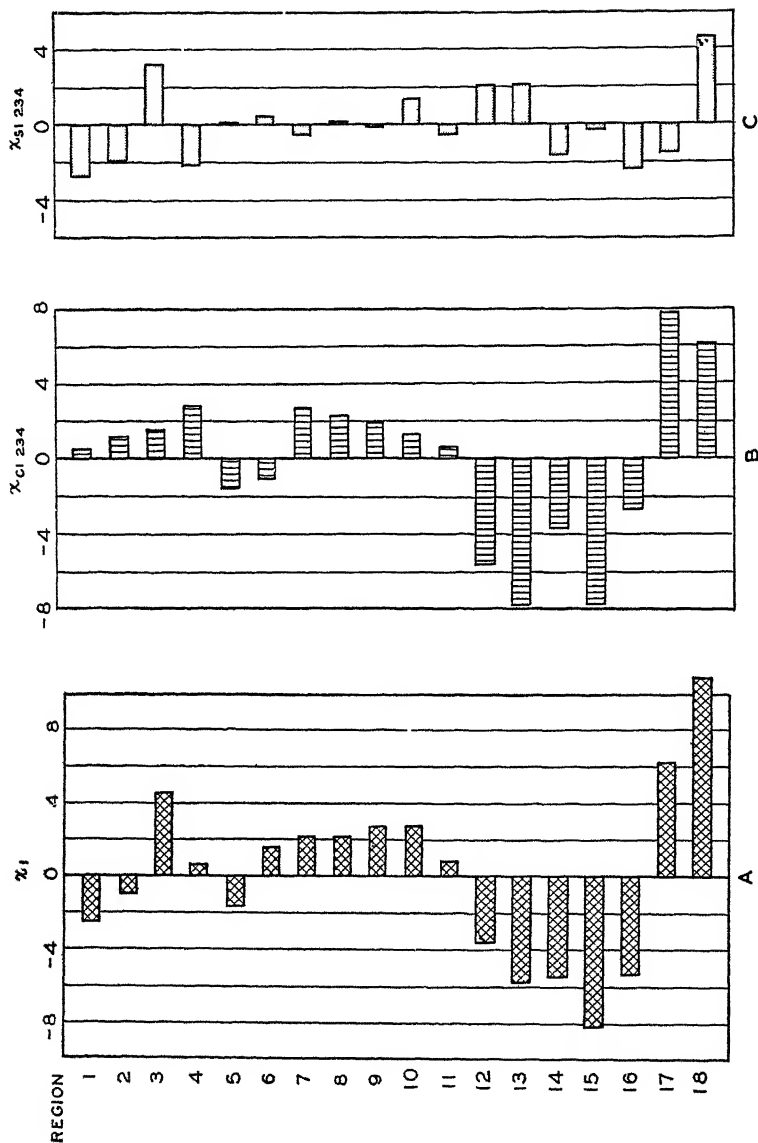
Chart 237. Deviations of Suicide Rates from Their Mean (x_1), Deviations from Their Mean of Computed Suicide Rates Based Upon Estimating Equation Using Average Age, Per Cent Male, and Business Failure Index as Independent Variables ($x_{C1 \cdot 234}$), and Deviations of Suicide Rates From Computed Rates ($x_{S1 \cdot 234}$). (Derived from data of Table 174.)

appears, on the whole, to have improved our estimates very little. Although the error in the New York City estimate is considerably reduced, several of the others have been increased somewhat, and the discrepancy in the case of California is nearly as great as before. Apparently the peculiar factors affecting the California suicide rate have not been included.

Three independent variables: partial correlation. Proceeding now in the usual fashion, we obtain partial correlation coefficients as follows:

$$r_{12.34}^2 = \frac{\Sigma x_{C1\ 234}^2 - \Sigma x_{C1\ 34}^2}{\Sigma x_1^2 - \Sigma x_{C1\ 34}^2} = \frac{313.044 - 221.179}{383.685 - 221.179} = \frac{91.865}{162.506} = .5653.$$

$$r_{12\ 34} = +.7519.$$

$$r_{13.24}^2 = \frac{\Sigma x_{C1\ 234}^2 - \Sigma x_{C1\ 24}^2}{\Sigma x_1^2 - \Sigma x_{C1\ 24}^2} = \frac{313.044 - 231.159}{383.685 - 231.159} = \frac{81.885}{152.526} = .5369.$$

$$r_{13.24} = +.7327.$$

$$r_{14.23}^2 = \frac{\Sigma x_{C1\ 234}^2 - \Sigma x_{C1\ 23}^2}{\Sigma x_1^2 - \Sigma x_{C1\ 23}^2} = \frac{313.044 - 302.217}{383.685 - 302.217} = \frac{10.827}{81.468} = .1329.$$

$$r_{14.23} = +.3646.$$

It might be thought that, as additional factors are held constant, the dependent variable would be progressively less closely associated with a given independent variable. For instance, the correlation between suicides X_1 and business failures X_4 was found to be $r_{14} = +.4040$; but, when the age factor X_2 was also brought into the picture (technically, when suicide rates and business failure index numbers were each adjusted for variations in average age), we had $r_{14.2} = +.0868$. What appeared to be a relationship between business failures and suicides was in fact largely a relationship between average age and suicide rates. On the other hand, $r_{13} = +.5925$ increased to $r_{13.2} = +.6856$ when age X_2 was taken into consideration. In this case the average age had varied in the different regions in such a way as to obscure the co-variation of per cent male X_3 and suicide rate X_1.

The reader should not necessarily conclude from these measures that differences in mental traits attributable to age and sex make for susceptibility to the urge of self-destruction. It may well be that older persons and males are more liable to find themselves confronted with situations leading to despondency. Thus financial worries have their first incidence on the chief breadwinner of the family, usually a mature male. Also, certain diseases of old age may partially account for the higher suicide rates among older persons. Whatever the conditions leading to suicide, it would appear that, taken together, they are fairly constant from region to region, but that they vary in their incidence with age and with

the proportion of males in the population.   There are some exceptions to this statement, notably California.   The introduction of more variables is needed to improve the accuracy of our estimates for this and some of the other regions, and so to increase the magnitude of our multiple coefficient of correlation.

Another Approach to Multiple and Partial Correlation

Partial coefficients.   The fact that partial correlation coefficients sometimes become higher and sometimes lower as more variables are held constant may be more clearly understood when we learn how the coefficients of higher order may be derived from those of lower order.   The values of the coefficients of *zero order* for the suicide study we have determined to be:

$$r_{12} = +.7742; r_{13} = +.5925; r_{14} = +.4040;$$
$$r_{23} = +.2183; r_{24} = +.4579;$$
$$r_{34} = -.1142.$$

As previously stated, they are called zero order coefficients because no variables are held constant.   From these zero order coefficients, *first order* coefficients, with one variable held constant, may be computed.   Below are given the formulae for the nine possible coefficients which may be computed for this problem, together with the substitutions and results (Three others, $r_{23\ 1}$, $r_{24\ 1}$, and $r_{34.1}$, have not been included, since they hold X_1 constant and do not concern this problem.)   Strictly speaking, only six of these coefficients are required for further computations—either the first six or the last six, although for checking purposes all nine may be desired.[5]

$$r_{12.4} = \frac{r_{12} - (r_{14})(r_{24})}{\sqrt{1 - r_{14}^2}\sqrt{1 - r_{24}^2}} = \frac{.7742 - (.4040)(.4579)}{(.9148)(.8890)} = +.7245;$$

$$r_{13\ 4} = \frac{r_{13} - (r_{14})(r_{34})}{\sqrt{1 - r_{14}^2}\sqrt{1 - r_{34}^2}} = \frac{.5925 - (.4040)(-.1142)}{(.9148)(.9935)} = +.7026;$$

$$r_{23\ 4} = \frac{r_{23} - (r_{24})(r_{34})}{\sqrt{1 - r_{24}^2}\sqrt{1 - r_{34}^2}} = \frac{.2183 - (.4579)(-.1142)}{(.8890)(.9935)} = +.3064;$$

$$r_{12.3} = \frac{r_{12} - (r_{13})(r_{23})}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}} = \frac{.7742 - (.5925)(.2813)}{(.8056)(.9759)} = +.8203;$$

$$r_{14.3} = \frac{r_{14} - (r_{13})(r_{34})}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{34}^2}} = \frac{.4040 - (.5925)(-.1142)}{(.8056)(.9935)} = +.5893;$$

---

[5] Proof that these formulae are the equivalent of those we have been using is given in Appendix B, section XXIV-2.   The labor of computation can be materially shortened if values of $\sqrt{1 - r^2}$ are looked up in J. R. Miner, *Tables of $\sqrt{1 - r^2}$ and $1 - r^2$ for Use in Partial Correlation and Trigonometry*, Johns Hopkins Press, Baltimore, 1922.

$$r_{24.3} = \frac{r_{24} - (r_{23})(r_{34})}{\sqrt{1 - r_{23}^2}\ \sqrt{1 - r_{34}^2}} = \frac{.4579 - (.2183)(-.1142)}{(.9759)(.9935)} = +.4979;$$

$$r_{13.2} = \frac{r_{13} - (r_{12})(r_{23})}{\sqrt{1 - r_{12}^2}\ \sqrt{1 - r_{23}^2}} = \frac{.5925 - (.7742)(.2183)}{(.6329)(.9759)} = +.6857;$$

$$r_{14.2} = \frac{r_{14} - (r_{12})(r_{24})}{\sqrt{1 - r_{12}^2}\ \sqrt{1 - r_{24}^2}} = \frac{.4040 - (.7742)(.4579)}{(.6329)(.8890)} = +.0880;$$

$$r_{34.2} = \frac{r_{34} - (r_{23})(r_{24})}{\sqrt{1 - r_{23}^2}\ \sqrt{1 - r_{24}^2}} = \frac{-.1142 - (.2183)(.4579)}{(.9759)(.8890)} = -.2469.$$

Six of these coefficients, those which correlate suicide rates with another variable, have previously been computed. Except for slight discrepancies in the fourth digit due to rounding, the two sets of results are in agreement.

For a four-variable problem there are three second order coefficients involving X_1 as the dependent variable. These may be computed:

$$r_{12.34} = \frac{r_{12.4} - (r_{13.4})(r_{23.4})}{\sqrt{1 - r_{13.4}^2}\ \sqrt{1 - r_{23.4}^2}} = \frac{.7245 - (.7026)(.3064)}{(.7115)(.9519)} = +.7518;$$

$$r_{13.24} = \frac{r_{13.4} - (r_{12.4})(r_{23.4})}{\sqrt{1 - r_{12.4}^2}\ \sqrt{1 - r_{23.4}^2}} = \frac{.7026 - (.7245)(.3064)}{(.6893)(.9519)} = +.7325;$$

$$r_{14.23} = \frac{r_{14.3} - (r_{12.3})(r_{24.3})}{\sqrt{1 - r_{12.3}^2}\ \sqrt{1 - r_{24.3}^2}} = \frac{.5893 - (.8203)(.4979)}{(.5719)(.8672)} = +.3647.$$

The above formulae employ the first six of the first order coefficients. If desired, the following formulae, using the last six coefficients, may be employed as a check.

$$r_{12.34} = \frac{r_{12.3} - (r_{14.3})(r_{24.3})}{\sqrt{1 - r_{14.3}^2}\ \sqrt{1 - r_{24.3}^2}} = \frac{.8023 - (.5893)(.4979)}{(.8079)(.8672)} = +.7521.$$

$$r_{13.24} = \frac{r_{13.2} - (r_{14.2})(r_{34.2})}{\sqrt{1 - r_{14.2}^2}\ \sqrt{1 - r_{34.2}^2}} = \frac{.6857 - (.0880)(-.2469)}{(.9961)(.9690)} = +.7329.$$

$$r_{14.23} = \frac{r_{14.2} - (r_{13.2})(r_{34.2})}{\sqrt{1 - r_{13.2}^2}\ \sqrt{1 - r_{34.2}^2}} = \frac{.0880 - (.6857)(-.2469)}{(.7279)(.9690)} = +.3648.$$

Again we find agreement, to three digits, with the same measures computed by the other method. If there are five variables, the formula for $r_{12.345}$ is

$$r_{12.345} = \frac{r_{12.45} - (r_{13.45})(r_{23.45})}{\sqrt{1 - r_{13.45}^2}\ \sqrt{1 - r_{23.45}^2}},\ \text{or}$$

$$= \frac{r_{12.35} - (r_{14.35})(r_{24.35})}{\sqrt{1 - r_{14.35}^2}\ \sqrt{1 - r_{24.35}^2}},\ \text{or}$$

$$= \frac{r_{12.34} - (r_{15.34})(r_{25.34})}{\sqrt{1 - r_{15.34}^2}\ \sqrt{1 - r_{25.34}^2}}.$$

If the reader has followed the exposition thus far, he will easily be able to construct formulae for the other third order coefficients, and also for those of higher order.

σ_S **and** R. It is possible now to obtain other measures of relationship by use of the different coefficients of partial correlation. For four variables the formulas and their application to the problem at hand are:[6]

$$\sigma_{S1\ 234}^2 = \frac{\Sigma x_1^2[(1 - r_{14}^2)(1 - r_{13\ 4}^2)(1 - r_{12\ 34}^2)]}{N}$$

$$= \frac{(383\ 65)(.8368)(.5063)(.4348)}{18} = \frac{(383.65)(.1842)}{18} = 3.926.$$

$\sigma_{S1.234} = 1.981.$

$R_{1.234}^2 = 1 - [(1 - r_{14}^2)(1 - r_{13\ 4}^2)(1 - r_{12.34}^2)] = 1 - .1842 = .8158.$

$R_{1\ 234} = .9032.$

It will be remembered from simple correlation that one approach considered r as the slope of the estimating line in terms of the standard deviation of the different variables. In terms of symbols used in this chapter,

$$r_{12} = b_{12} \div \frac{\sigma_1}{\sigma_2}, \text{ and } b_{12} = r_{12} \frac{\sigma_1}{\sigma_2}.$$

But

$$\sigma_1 = \frac{\sigma_{S1\ 2}}{\sqrt{1 - r_{12}^2}}, \quad \sigma_2 = \frac{\sigma_{S2\ 1}}{\sqrt{1 - r_{21}^2}}, \text{ and } r_{12}^2 = r_{21}^2.$$

Therefore

$$\frac{\sigma_1}{\sigma_2} = \frac{\sigma_{S1\ 2}}{\sigma_{S2.1}},$$

and substituting we find that

$$r_{12} = b_{12} \div \frac{\sigma_{S1\ 2}}{\sigma_{S2.1}}, \text{ and } b_{12} = r_{12} \frac{\sigma_{S1.2}}{\sigma_{S2.1}}.$$

By analogy, then, we have

$$r_{12.34} = b_{12.34} \div \frac{\sigma_{S1\ 234}}{\sigma_{S2.134}}, \text{ and } b_{12\ 34} = r_{12.34} \frac{\sigma_{S1.234}}{\sigma_{S2.134}}.$$

To obtain $\sigma_{S2.134}$, it is convenient to use the formula

$$\sigma_{S2.134}^2 = \frac{\Sigma x_2^2(1 - r_{24}^2)(1 - r_{23\ 4}^2)(1 - r_{21\ 34}^2)}{N},$$

$$= \frac{(61.4119)(.7903)(.9061)(.4348)}{18} = 1.062.$$

$\sigma_{S2\ 134} = 1.031.$

_____
[6] For an explanation of these formulae, see Appendix B, section XXIV–3.

Proceeding, we find that

$$b_{12.34} = .7518 \frac{1.981}{1.031} = 1.444.$$

As in our earlier illustrations, there is a slight discrepancy in the fourth digit between this and the previously described method, due to rounding.   The other coefficients of estimation may be obtained by substituting in the formulae

$$b_{13.24} = r_{13.24} \frac{\sigma_{S1.234}}{\sigma_{S3.124}};$$

$$b_{14.23} = r_{14\,23} \frac{\sigma_{S1.234}}{\sigma_{S4.123}}.$$

Although several variations of the formulae for the different standard errors are possible, the following are convenient in that they require only the correlation coefficients given on pages 770–771:

$$\sigma^2_{S3.124} = \frac{\Sigma x_3^2 (1 - r_{34}^2)(1 - r_{23\,4}^2)(1 - r_{13\,24}^2)}{N};$$

$$\sigma^2_{S4\,123} = \frac{\Sigma x_4^2 (1 - r_{34}^2)(1 - r_{24\,3}^2)(1 - r_{14\,23}^2)}{N}.$$

Other Measures of the Individual Importance of the Independent Variables

It will be recognized that the methods of obtaining partial correlation coefficients which have been described are very laborious, since they necessitate either the solution of three extra sets of normal equations that have no function other than to obtain the values $\Sigma x_{c1.23}^2$, $\Sigma x_{c1\,24}^2$, and $\Sigma x_{c1\,34}^2$, or the building up of various simple and partial correlation coefficients which likewise may be of no direct interest.   Consequently other measures of the importance of the individual factors are frequently used instead, which are much easier to compute.

Perhaps the most common of these are the *beta coefficients*:

$$\beta_{12.34}, \ \beta_{13.24}, \text{ and } \beta_{14\,23}.$$

The reader should not confuse these measures with β_1 and β_2, which were used to describe a frequency distribution.   The two sets of measures are entirely different in nature.   It will be recalled that the following relationship obtained in simple correlation:

$$r_{12} = b_{12} \frac{\sigma_2}{\sigma_1}.$$

Reasoning by analogy, we have:

$$\beta_{12.34} = b_{12.34} \frac{\sigma_2}{\sigma_1};$$

$$\beta_{13.24} = b_{13.24} \frac{\sigma_3}{\sigma_1};$$

$$\beta_{14.23} = b_{14.23} \frac{\sigma_4}{\sigma_1}.$$

The analogy is imperfect, since the standard deviations used with the *net* coefficient of estimation are *gross* measures of dispersion; that is, the variables have not been adjusted for variations in the other factors which have been held constant statistically.   The four standard deviations are readily found   They are: $\sigma_1 = 4\ 617$; $\sigma_2 = 1.847$, $\sigma_3 = .9284$; $\sigma_4 = 33.20$.   Substituting in the above formulae, we find:

$$\beta_{12\ 34} = 1.445\ \frac{1.847}{4.617} = +.578;$$

$$\beta_{13\ 24} = 2.429\ \frac{.9284}{4.617} = +.488;$$

$$\beta_{14.23} = .02711\ \frac{33.20}{4.617} = +.195.$$

The rank of the β coefficients in this case is the same as the partial coefficients.   This will usually, though not always, be the case.   Per cent male seems somewhat less important by this method, however

Two other measures of individual importance are sometimes used.   The *coefficients of separate determination* split up the expression for $R^2_{1.234}$ as used on page 766; hence we have

$$R^2_{1\ 234} = \frac{\Sigma x^2_{c1\ 234}}{\Sigma x^2_1} = \frac{b_{12\ 34}\Sigma x_1 x_2 + b_{13\ 24}\Sigma x_1 x_3 + b_{14\ 23}\Sigma x_1 x_4}{\Sigma x^2_1},$$

split into three components:

$$d^2_{12.34} = \frac{b_{12\ 34}\Sigma x_1 x_2}{\Sigma x^2_1}; \; d^2_{13\ 24} = \frac{b_{13.24}\Sigma x_1 x_3}{\Sigma x^2_1}; \; d^2_{14.23} = \frac{b_{14\ 23}\Sigma x_1 x_4}{\Sigma x^2_1}$$

The sum of three coefficients of separate determination, therefore, equals the coefficient of multiple determination.   These separate coefficients, however, are thought to be more subject to random error than the β coefficients; furthermore, each includes *part* of the joint determination of the other two independent variables.   Another disadvantage of this coefficient is that the value of d^2 may be negative and thus a coefficient of *separate* correlation d cannot be obtained.   (See Mordecai Ezekiel, *Methods of Correlation Analysis*, pp 380–383, John Wiley and Sons, New York, 1930.)   Another measure of individual importance, not widely used as yet but which Ezekiel recommends, is the *coefficient of part correlation*.   This coefficient measures the correlation between an independent variable and the dependent variable, the latter only having been adjusted for net variations in the other independent variables.   Perhaps the relationship between multiple, partial, and part correlation will be clearer if we think of them as follows (in terms of a 4-variable problem):

Multiple correlation. $R_{1.234}$ = simple correlation of

$$X_1 \text{ with } X_{c1\ 234}$$

Partial correlation: $r_{12.34}$ = simple correlation of

$$[X_2 - b_{23.4}\ X_3 - b_{24.3}X_4] \text{ with } [X_1 - b_{13.4}\ X_3 - b_{14.3}X_4].$$

Part correlation: $_{12}r_{34}$ = simple correlation of

$$X_2 \text{ with } [X_1 - b_{13.24}\ X_3 - b_{14.23}\ X_4].$$

See also Ezekiel, *ibid.*, pp. 181–183.

Estimate of Correlation in the Population

As is the case with simple linear or non-linear, so with multiple or partial correlation it may be desirable to estimate the correlation that exists in the population.   The formula to use for multiple correlation is identical with the one with which we are familiar:

$$\bar{R}^2_{1\ 234\ \ldots m} = \frac{R^2_{1\ 234\ \ldots m}\,(N-1) - (m-1)}{N-m},$$

where m is the number of degrees of freedom lost; that is, the number of constants in the estimating equation, including $a_{1\ 234}\ \ldots$.   In the present instance we have for the multiple correlation coefficient

$$\bar{R}^2_{1\ 234} = \frac{.8159(18-1) - (4-1)}{18-4} = .7764.$$
$$\bar{R}_{1\ 234} = .8811.$$

For the partial correlation coefficient ($\bar{r}_{14\ 23}\ \ldots$, for example), we have a slightly different expression:[7]

$$\bar{r}^2_{14\ 23\ \ldots} = \frac{r^2_{14.23}\ \ldots (N-m+1) - 1}{N-m}.$$

Applying this formula, we have

$$\bar{r}^2_{14.23} = \frac{.1329(18-4+1) - 1}{18-4} = .0710.$$
$$\bar{r}_{14\ 23} = +.2665.$$

Reliability of Coefficients

Standard error of coefficients.   Measures sometimes used to test the reliability of multiple and partial correlation coefficients are analogous to the formula for σ_r employing the coefficient obtained from the sample,

$$\sigma_r = \frac{1-r^2}{\sqrt{N-m}}.$$

For multiple correlation this may be stated

$$\sigma_{R_{1.234}\ \ldots m} = \frac{1 - R^2_{1\ 234\ \ldots m}}{\sqrt{N-m}};$$

---

[7] This expression develops from the relationship

$$1 - \bar{r}^2_{14.23} = \frac{1 - \bar{R}^2_{1\ 234}}{1 - \bar{R}^2_{1.23}},$$

as shown in Appendix B, section XXIV–4.

and for partial correlation[8]

$$\sigma_{r_{12\ 34\ \ldots\ m}} = \frac{1 - r^2_{12.34\ \ldots\ m}}{\sqrt{N - m}}.$$

As previously noted on page 681, such expressions are grossly inaccurate when N is small, or when the value of r is large. An additional limitation to σ_R is that the sampling distribution of R varies with the magnitude of m. It is therefore preferable to make use of more exact methods, such as the analysis of variance.

Analysis of variance. Let us summarize in tabular form some of the major results of our correlation analysis. All the needed data concerning variation will be found from the following expressions, which were used in computing various correlation coefficients and all of which have been given on preceding pages:

$$r^2_{12} = \frac{229.971}{383.685} = .5994 \qquad r_{12} = +.7742$$

$$r^2_{13\ 2} = \frac{72.246}{153.714} = .4700 \qquad r_{13.2} = +.6856$$

$$R^2_{1\ 23} = \frac{302.217}{383.685} = .7877 \qquad R_{1\ 23} = .8875$$

$$r^2_{14\ 23} = \frac{10.827}{81.468} = .1329 \qquad r_{14.23} = +.3646$$

$$R^2_{1\ 234} = \frac{313\ 044}{383.685} = .8159 \qquad R_{1.234} = .9033$$

Source of variation in X_1	Amount of variation	Degrees of freedom	Variance
Gross amount explained by X_2... .	229 971	1	229 971
Increment explained by addition of X_3	72 246	1	72 246
Total explained by X_2 and X_3.	302 217	2	151 108
Increment explained by addition of X_4	10.827	1	10 827
Total explained by X_2, X_3, and X_4.	313 044	3	104.348
Residual, unexplained by X_2, X_3, and X_4	70 641	14	5.046
Total.	383 685	17	22.570

[8] If the population value of $r_{12.34\ \ldots\ m}$ is used in the formula, this expression is

$$\sigma_{r_{12\ 34\ \ldots\ m}} = \frac{1 - r^2_{P12.34\ \ldots\ m}}{\sqrt{N - m + 1}}.$$

The distribution of such sample coefficients may be considered to approximate normality only when the population coefficient is small and N is large.

We may now test the significance of $R_{1\,234}$ by the use of

$$F = \frac{104\,348}{5.046} = 20.679.$$

Appendix G2 indicates that for the .001 level of significance, when $n_1 = 3$ and $n_2 = 14$, F should equal 9.730. The results indicate that $R_{1.234}$ is significant.

If we wish to test $r_{13\,2}$, it is best to relate the increment in variance attributable to variable 3 to the variance which is attributable to chance, the latter being the variance not accounted for by *all* the independent variables taken together. Thus we have

$$F = \frac{72.246}{5.046} = 14.317.$$

When $n_1 = 1$ and $n_2 = 14$, F should equal 8.862 for the .01 level of significance, and 17.143 for the .001 level. Clearly $r_{13\,2}$ is significant. Notice that the unexplained variance is derived from 70.641 rather than from $383.685 - 302.217 = 81.468$. The latter quantity is not the variation due to chance factors, but is the variation due to chance factors plus variable 4.

If we had not made use of variable 4 in our correlation analysis, we should have used, for the unexplained variance, $5.4313 = 81.468 \div 15$. We should then have computed

$$F = \frac{72.246}{5.4313} = 13.302.$$

We now have a smaller value for F; however, this is partially offset by the fact that n_2 is 15 instead of 14, and F need not be so high for the same level of significance. In the present case it is not obvious whether it is more accurate to use for the unexplained variance that which remains after employing variables 2, 3, and 4, or that which remains after using variables 2 and 3 only, since the F test, as illustrated in the following paragraph, fails to show that the additional explanation attributed to variable 4 is significant. In general, however, the test using fewer independent variables is not so accurate as the test using more, and the former may erroneously fail to show significance for the factor being tested.

To test the significance of $r_{14.23}$, we compute

$$F = \frac{10.827}{5\,046} = 2.146.$$

The F table for $P = .05$, when $n_1 = 1$ and $n_2 = 14$, requires that $F = 4.600$. Therefore, $r_{14.23}$ cannot be regarded as significant.

The significance of a partial correlation coefficient may also be tested

by use of the t table, and the conclusions from such a test agree with the analysis of variance test. We compute

$$t = \frac{r_{12\ 3\ .\ \ m}\sqrt{N - m}}{\sqrt{1 - r^2_{12\ 3\ \ldots\ m}}}.$$

Using the t test to ascertain the significance of $r_{13.2}$, we have

$$t = \frac{.6858\sqrt{18 - 3}}{\sqrt{1 - .4700}} = 3.648,$$

which, according to the t table, lies beyond the .01 level of significance and is in agreement with the F test. It is interesting to note that the value of t is the square root of the value of F, when $n_1 = 1$.

We may also transform a partial correlation coefficient into Z in order to discover if the coefficient differs significantly from some known or hypothetical population value, or from some other observed correlation coefficient. The general formula is

$$Z = 1.15129 \log_{10} \frac{1 + r}{1 - r},$$

with standard error

$$Z = \frac{1}{\sqrt{N - m - 1}}.$$

Multiple Curvilinear Correlation

Transformation to linear form. As was found true with gross relationships, so the net relationship between a dependent variable and one or more independent variables is sometimes non-linear. Sometimes it is possible to reduce such non-linear relationships to linear form by using logarithms or reciprocals (or possibly some other function) of one or more variables. Thus, with three variables, we might have an estimating equation of one of the following types:

$$X_{C1.23} = a_{1.23} + b_{12.3}\log X_2 + b_{13\ 2}X_3;$$
$$X_{C1.23} = a_{1.23} + b_{12.3}\log X_2 + b_{13\ 2}\sqrt{X_3};$$
$$X_{C1\ 23} = a_{1.23} + b_{12.3}\log X_2 + b_{13.2}\frac{1}{X_3}.$$
$$\log X_{C1.23} = \log a_{1.23} + X_2 \log b_{12.3} + \frac{1}{X_3}\log b_{13.2};$$
$$\log X_{C1.23} = \log a_{1.23} + X_2 \log b_{12\ 3} + X_3 \log b_{13.2};$$
$$\log \log X_{C1\ 23} = \log a_{1.23} + X_2 \log b_{12.3} + X_3 \log b_{13.2}.$$

The above types are, of course, but six of a number of possible combinations, a proper choice among which should, in perhaps a majority of cases, result in empirical curves satisfactory for purposes of estimation. It

would not, however, be possible to transform into a linear equation a relationship in which the logs of the dependent variable were related to the second variable and the reciprocals of the dependent variable were related to the third variable.

Use of polynomials. Even more flexible is the use of polynomials, and probably more useful as an exploratory tool in that it is not necessary to have a very precise hypothesis concerning the nature of the relationship among the variables before undertaking the correlation. Thus we might, using three variables, start with the equation type

$$X_{C1\ 23} = a_{1\ 23} + b_{12.3}X_2 + b_{13\ 2}X_3;$$

then using degrees of freedom compute estimates of explained variance $\dfrac{\Sigma x^2_{C1\ 23}}{2}$, and of unexplained variance $\dfrac{\Sigma x^2_{S1\ 23}}{N-3}$, and test for significance, using the F or z table.

Then we could include also the squares of X_2 in our equation, in this fashion:

$$X_{C1.22'3} = a_{1.22'3} + b_{12\ 2'3}X_2 + b_{12'\ 23}X_2^2 + b_{13\ 22'}X_3.$$

To test whether the use of the second powers of X_2 has significantly reduced the variance, we should relate the increase in the explained variance $\dfrac{\Sigma x^2_{C1\ 22'3}}{3} - \dfrac{\Sigma x^2_{C1\ 23}}{2}$ to the variance that is still unexplained $\dfrac{\Sigma x^2_{S1\ 22'3}}{N-4}$. If the test indicates that the reduction in variance is significant, we should conclude that it is worth while to use the additional constant $b_{12'.23}$. In similar fashion we could utilize the squares of X_3, or higher powers of both X_2 and X_3.

Of course, it is not always necessary to go through all the labor suggested in the preceding paragraph. The statistician can frequently decide on economic or other non-mathematical grounds the type of relationship which exists among the variables. Thus economists are of the opinion that demand curves usually slope downward to the right, and are concave upward. This would indicate that the second powers of the price series should be used if the response of purchasers to various prices are to be estimated by the use of simple polynomials. In the illustration that follows, millings of wheat (X_1) are to be correlated with the price of flour (X_2) and index numbers of income of industrial workers (X_3). The data were assembled by the Bureau of Agricultural Economics of the United States Department of Agriculture;[9] and the net relationship between X_1 and

---

[9] Instead of adopting the usual practice, when correlating time series, of expressing the data as percentages of trend, we are following the procedure used by the Bureau of Agricultural Economics in analyzing this problem. The Bureau did not adjust for trend, apparently on the assumption that the trends were approximately horizontal, and that the period was too short to permit of accurate trend measurement.

X_2, with X_3 held constant, was estimated by this Bureau in order to aid in determining the effect of a processing tax on wheat acreage. The presumption was that, if it could be shown that the demand for flour, and hence for wheat, by millers was inelastic (only slightly affected by the price of flour), farmers would not find it necessary to reduce materially their wheat acreage.

TABLE 176

WHEAT MILLED, PRICE OF FLOUR, AND INDEX NUMBERS OF INCOME OF INDUSTRIAL WORKERS, BY YEARS, 1924–1935

Year beginning July 1	Wheat milled (millions of bushels) X_1	Average price of flour per barrel (dollars) X_2	Income of industrial workers (1924–1929 = 100) X_3
1924	475	7.94	94 0
1925	490	8 36	100.5
1926	493	7.42	101.8
1927	494	7 36	98.5
1928	500	6.29	102 6
1929	496	6.48	100.8
1930	481	4.78	77.3
1931	474	3.84	56.8
1932	481	3.86	40 6
1933	435	6.47	54.7
1934	443	6.66	60.9
1935	460	6.78	68 8
Total	5,722	76.24	957.3

Source United States Department of Agriculture, Bureau of Agricultural Economics   Published in pamphlet of United States Treasury Department, *An Analysis of the Effect of a Processing Taxes Levied Under the Agricultural Adjustment Act, 1 37, p 8*   Wheat milled is that milled for domestic consumption   Average price is a simple average of winter wheat straights Kansas City, and spring wheat family patents, Minneapolis   Index numbers of income are for year beginning June 1.

The data are shown in Table 176, and the gross relationship between X_1 and X_2, and between X_1 and X_3, may be noted by reference to Chart 238. Although the milling of wheat seems to be directly related to income of industrial workers, and the relationship is apparently linear, scarcely any relationship is discernible between wheat milled and the price of flour. But using our knowledge of the usual shape of demand curves, we may hypothecate the following type of relationship:

$$X_{C1.22'3} = a_{1.22'3} + b_{12.2'3}X_2 + b_{12'.23}X_2^2 + b_{13.22'}X_3.$$

The normal equations required are:

I. $\Sigma X_1 = N a_{1\,22'3} + b_{12\,2'3}\Sigma X_2 + b_{12'.23}\Sigma X_2^2 + b_{13\,22'}\Sigma X_3;$

II. $\Sigma X_1 X_2 = a_{1.22'3}\Sigma X_2 + b_{12\,2'3}\Sigma X_2^2 + b_{12'.23}\Sigma X_2^3 + b_{13\,22'}\Sigma X_2 X_3;$

III. $\Sigma X_1 X_2^2 = a_{1\,22'3}\Sigma X_2^2 + b_{12\,2'3}\Sigma X_2^3 + b_{12'\,23}\Sigma X_2^4 + b_{13.22'}\Sigma X_2^2 X_3;$

IV. $\Sigma X_1 X_3 = a_{1\,22'3}\Sigma X_3 + b_{12\,2'3}\Sigma X_2 X_3 + b_{12'.23}\Sigma X_2^2 X_3 + b_{13.22'}\Sigma X_3^2.$

A computation table of the various sums and product sums will not be shown, since no new principle is involved, but on page 782 is shown a check on the accuracy of the computations. To find a particular product sum, we locate the multiplicand in the stub and read across to the ap-
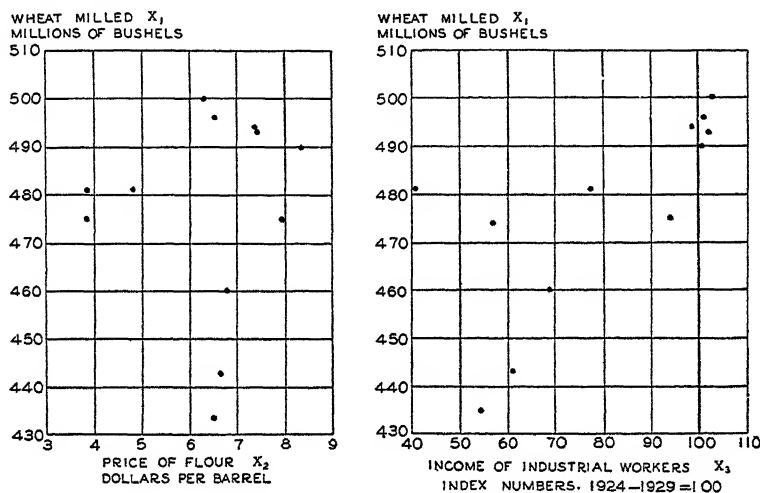
Chart 238.   Scatter Diagrams of Gross Relationship of Price of Flour and Income of Industrial Workers with Amount of Wheat Milled, by Years, 1924-1935.   (Data of Table 176.)

propriate multiplier column.   Thus it will be seen that the value of $\Sigma X_2^2 X_3$, which is 43,544.6876, is found at the intersection of row X_2^2 and column X_3.   The dotted lines indicate the items that are totaled to check with the X_Σ column.

From the computations on page 782 the normal equations are formed:

I. $5,722 = 12 a_{1\,22'3} + 76.24 b_{12\,2'3} + 508\,3922 b_{12'.23} + 957\,3 b_{13.22'}.$

II. $36,380\,51 = 76\,24 a_{1\,22'3} + 508.3922 b_{12.2'3} + 3,514.2712 b_{12'\,23} + 6,335\,683 b_{13.22'}$

III $242,854\,1787 = 508\,3922 a_{1.22'3} + 3,514\,2712 b_{12\,2'3} + 24,947\,444 b_{12'.23} + 43,544\,6876 b_{13\,22'}.$

IV. $460,092\,5 = 957.3 a_{1\,22'3} + 6,335\,683 b_{12\,2'3} + 43,544.6876 b_{12'\,23} + 81,973.37 b_{12'\,22'}.$

Solving the normal equations simultaneously gives the estimating equation:

$$X_{C1.22'3} = 558.4216 - 48.34186 X_2 + 3.145498 X_2^2 + 1.156772 X_3.$$

Entries in the boxes below are sums or sums of products as indicated by column and row headings.

	X_1	X_2	X_2^2	X_3	X_Σ
1	5,722.	76.24	508.3922	957.3	7,263.9322
X_1	2,733,298.	36,380.51	242,854.1787	460,092.5	3,472,625.18
X_2	...	508.3922	3,514.2712	6,335.683	46,738.8564
X_2^2	...	...	24,947.4444	43,544.6876	314,860.580
X_3	...	...	...	81,973.37	591,946.241

The other measures of relationship are now easily found in the usual fashion. For the explained sums of squares we have

$$\Sigma X^2_{C1.22'3} = a_{1.22'3}\Sigma X_1 + b_{12.2'3}\Sigma X_1 X_2 + b_{12'\ 23}\Sigma X_1 X^2_2 + b_{13\ 22'}\Sigma X_1 X_3$$
$$= 2,732,705.$$

The total sums of squares is $\Sigma X^2_1 = 2,733,298$. The usual correction factor is $\overline{X}_1\Sigma X_1 = 2,728,440$. From these values we may compute the different measures of variation:

Total	Σx^2_1	$= \Sigma X^2_1$	$- \overline{X}_1\Sigma X_1$	$= 2,733,298 - 2,728,440 =$	4,858
Explained	$\Sigma x^2_{C1\ 22'3} = \Sigma X^2_{C1\ 22'3} - \overline{X}_1\Sigma X_1$			$= 2,732,705 - 2,728,440 =$	4,265
Unexplained.	$\Sigma x^2_{S1\ 22'3} = \Sigma X^2_1$	$- \Sigma X^2_{C1.22'3} =$		593	$= 593$

We therefore have

$$\sigma^2_{S1\ 22'3} = \frac{\Sigma x^2_{S1\ 22'3}}{N} = \frac{593}{12} = 49.52, \text{ and } \sigma_{S1\ 22'3} = 7.03.$$

$$R^2_{1\ 22'3} = \frac{\Sigma x^2_{C1\ 22'3}}{\Sigma x^2_1} = \frac{4,265}{4,858} = .8879, \text{ and } R_{1\ 22'3} = .937.$$

$R_{1\ 23}$, computation of which is not shown, is .889. Although the nonlinear correlation is very high, it must be remembered that we had only twelve observations and our four constants, $a_{1\ 22'3}$, $b_{12.2'3}$, $b_{12'.23}$, and $b_{13.22'}$, have used up four degrees of freedom, so that we have only eight degrees of freedom left. Furthermore, Chart 239 shows that the mathematical equation of the *net* relationship between the price of flour and the amount of wheat milled, when income of industrial workers is held constant, is not strictly logical, for the solid curve on this chart turns up after a price of about $7.50 per bushel is reached. The equation of this curve of net relationship is

$$X_{C1.22'(3)} = 650.704 - 48.34186X_2 + 3.145498X^2_2,$$

in which $X_{C1\ 22'(3)}$ refers to a value of variable X_1 (wheat millings) as estimated from the curvilinear relationship with variable X_2 (price of flour), after allowing for the effect of variations in X_3 (income of industrial workers). The symbol (3) indicates that X_3 has been held constant statistically. The net estimating equation was obtained by substituting 79.775 (i.e., $\overline{X}_3$) for X_3 in our original estimating equation. The broken line on this chart, which represents a more logical relationship, was obtained by a graphic process which will be described in the final section of this chapter.

Before leaving this section, however, it is worth noting that it is possible to use reciprocals or logarithms of some of the variables and at the same time utilize the higher powers of any or all of the independent variables, whether they are in original form, or transformed into reciprocals or loga-

rithms. This possibility, of course, materially increases the flexibility of this approach.

A graphic approach. Statisticians in the United States Department of Agriculture have developed an extremely flexible technique by which curves of net relationship and a coefficient of multiple correlation may be obtained through successive approximations by means of charts and mathematics no more advanced than simple arithmetic. While this method has distinct limitations, it is useful as an exploratory tool in determining the appropriate type of equation to fit by mathematical methods.

Chart 240 contains ordinary scatter diagrams of the relationship between price of flour and amount of wheat milled, but with additional information included also. Remembering that a curve of net relationship is supposed to represent the relationship between two variables when certain other variables are held constant, we can obtain a first approximation to the net relationship between the milling of wheat and the price of flour by proceeding as follows: First, we note from Table 176 that in the years 1925–1929 inclusive, the income of industrial workers was substantially constant. Those years are indicated on section A of Chart 240 by black triangles. Through these triangles a broken line is fitted freehand. The level of industrial income, though by no means constant between 1930 and 1935, was at a distinctly lower level than for the earlier period. Consequently, these observations are shown by black dots, and a second broken line is fitted freehand to them. The year 1924 was intermediate between these groups with respect to industrial income, and is shown by a hollow triangle. Had there been other years falling naturally into a group with 1924, we should have fitted a third line. Using our two broken lines as guides, we have drawn in the solid curved line, which is our first approximation (line I) of the net functional relationship between X_1 and X_2. The reason the line is drawn as a curve is partly because this is the type of relationship we should logically expect, but mainly because our chart has

WHEAT MILLED X_1
MILLIONS OF BUSHELS

MATHEMATICAL CURVE

GRAPHIC APPROXIMATION

PRICE OF FLOUR X_2
DOLLARS PER BARREL

Chart 239. Least Squares Estimate and Graphic Estimate of Net Relationship Between Price of Flour and Amount of Wheat Milled. (The dots are placed on the chart to represent deviations from the mathematically fitted curve. Derived from data of Table 176.)

two broken lines, the steeper one passing through the dots (which are mainly on the left-hand side of the chart) and the flatter one passing through the triangles (which are mainly on the right-hand side of the chart). It is not necessary that the first approximation be a good fit. Successive approximations, if made with skill, will correct the original discrepancies; however, the more accurate the first approximation is made, the less laborious is the completion of the procedure.

A slightly different technique is shown in section B of Chart 240. It is not always possible to divide the data into natural groups; but we can connect dots representing the different observations in the sequence of
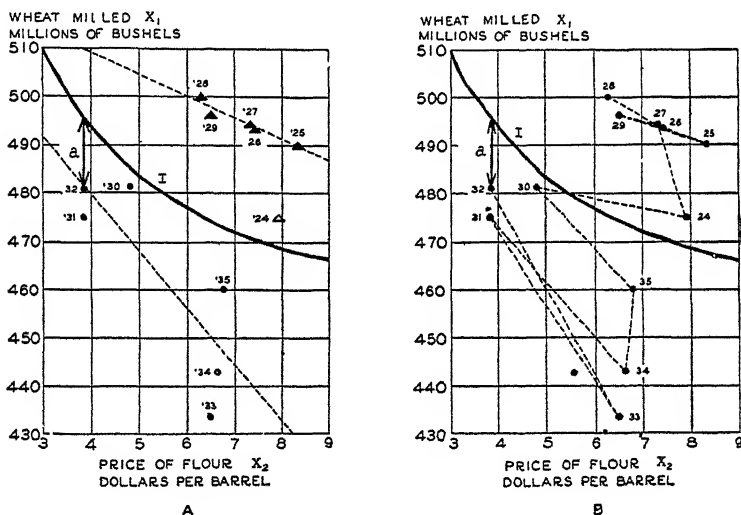
Chart 240. First Approximation to Net Non-Linear Relationship Between Price of Flour (X_2) and Amount of Wheat Milled (X_1), as Obtained by Two Different Techniques.

their rank with respect to the third variable. Thus the various years rank from lowest to highest, with respect to income of industrial workers, as follows: 1932; 1933; 1931; 1934; 1935; 1930; 1924; 1927; 1925; 1929; 1926; 1928. Dots representing those years have been connected in that order by broken lines. Wherever large gaps occur with respect to values of X_3, the broken lines may be omitted. Thus we might not have connected 1924 with either 1930 or 1926. The broken lines in this chart are intended as a guide to drawing in the first approximation line. With the exception of the lines connecting 1934 and 1935, and 1924 and 1927, the picture is clear. The solid line (I) is identical with that of section A Had there been four variables, we should have connected by dotted lines

only those observations which were approximately the same with respect to the variables X_3 and X_4. A scatter diagram of these two variables would have aided in locating such observations. Had there been also a fifth variable, the number of preliminary guide lines would have been restricted to observations similar with respect to X_3, X_4, and X_5. As the number of variables is increased, there are fewer and fewer dots that can be connected, and these dots are more and more difficult to find.

Still another method of obtaining a first approximation line is to use the equation obtained by the ordinary mathematical multiple correlation approach.

Chart 241. First Approximation to Net Relationship Between Income of Industrial Workers (X_3) and Deviations from the Mean in Amount of Wheat Milled. (Vertical deviations from zero on this chart are deviations from curve I of Chart 240 A or B)

The vertical distances between the different observations and the solid line (I) of Chart 240 represent the individual variations in X_1 that have not been explained by the different values of X_2. These values are now read from the chart and plotted against the values of X_3. Thus line a of Chart 241 is the same length as line a of Chart 240. It is apparent that a straight line fits these residuals very well, except for the 1932 observation. A straight line is therefore drawn in by inspection as a first approximation (line II) of the relationship between wheat milled (adjusted for the price of flour) and income of industrial workers.

Deviations from the curve of Chart 241 are now plotted above and below the first approximation curve (I) of the net relationship between X_1 and X_2. This is done on Chart 242; the first approximation curve (I) is shown here as a thin solid line. Note that line b of Chart 242 is the same length as line b of Chart 241. It is apparent that our first approximation curve (I) needs to be corrected. The heavy line represents a second approximation (I').

A further (and in this case, final) step is to plot deviations from the second approximation curve (I') of the net relationship of X_1 to X_2 above and below the first approximation curve (II) of the net relationship between wheat milled and income of industrial workers. This is done in Chart 243. Note this time that line c of Chart 243 is the same length as

line c of Chart 242. Although the 1932 observation is still too far off, there is not adequate reason for changing the shape of the first approximation curve, and so II' is the same as II. Had this curve been changed, it would have been necessary to proceed to obtain third approximation curves. It should be noted that the scatter about the estimating lines becomes smaller and smaller as the estimated effect of additional variables is removed, or as successive approximations of any relationship are made. The scatter can, of course, be reduced to zero if extremely complex curves are used; how far to go is a matter of judgment.

The final lines of relationship in these charts are not least square fits. Such lines cannot be obtained by purely graphic methods, although it is possible to so draw the curves on each chart that the deviations will total zero. The additional labor involved in adjusting the curves to do this exactly is probably not justified; it can be done by inspection with sufficient accuracy for most purposes. Reference to Chart 239 indicates that the average height of the curves determined by the two methods, mathematical and graphical, is about the same, and the patterns of the two are in substantial agreement, except that the mathematical curve turns up at the right.

Milling estimates for the different years must be computed by the addition of readings from curves I' and II'. The tabulation for 1924 will be:

Chart 242. Second Approximation to Net Relationship Between Price of Flour (X_2) and Amount of Wheat Milled (X_1). (Deviations from curve II are plotted around curve I.)

	X_1
Source	Curve reading
Curve I' (Chart 242) . .	464 9
Curve II' (Chart 243). . .	15 7
Estimate	480 6

Since the wheat milled in 1924 actually was 475.0, the residual is $475.0 - 480.6 = -5.6$ for that year. Unexplained variations for the other years can be computed in similar fashion, or they may be read directly from either of the final approximation charts (Chart 242 or 243). The standard error of estimate may now be obtained by squaring each deviation ($x_{\text{S1.2'3}}$), summing the squares, dividing by N, and extracting

the square root. (The subscript $2'$ here indicates an unspecified non-linear relationship with X_2.)

The coefficient of curvilinear correlation can easily be computed thus:

$$R^2_{1\,2'3} = 1 - \frac{\Sigma x^2_{81\,2'3}}{\Sigma x^2} = 1 - \frac{546.28}{4,858} = .8876.$$

$$R_{1\,2'3} = .942.$$

Graphic methods have not been devised for obtaining measures strictly analogous to coefficients of partial correlation. Nevertheless, much can be learned concerning the relative importance of the different variables simply from inspection of the final charts. Roughly, it may be said that the importance of the different factors varies directly with the vertical distance occupied on the charts by the different curves.

WHEAT MILLED X_1
MILLIONS OF BUSHELS

INCOME OF INDUSTRIAL WORKERS X_3

Chart 243. Income of Industrial Workers (X_3) and Deviations of Wheat Milled From Curve I' Plotted Around Curve II. (No second approximation to the net relationship between income of industrial workers and variations in amount of wheat milled has been made.)

Limitations of graphic method. Although the graphic method is extremely flexible, it is also highly subjective. Rarely would two statisticians obtain curves exactly alike from the same data. Consequently, good results can be obtained only by persons of experience and good judgment. This is in contrast with the mathematical procedure based on the method of least squares, in which case any competent computer can obtain only one possible result for a given equation type. A practical difficulty also is inherent in the method when a large number of variables are employed. The shape of each curve is determined not solely by the appearance of the individual scatter diagram; in drawing any first approximation curve, special importance is accorded to those observations which remain constant with respect to the other variables, and, when making successive approximations, consideration must be given to the effect which will be had on other scatter diagrams based on the residuals from the chart in question. Effective technique for so doing is lacking when there are more than three independent variables. It must also be remembered that, as more bends are introduced in the estimating curves, additional degrees of freedom are lost and the results become less reliable. Nor is it possible to

say exactly how many constants are involved in a freehand curve; consequently the reliability of the coefficients can be appraised only roughly by estimating the number of constants involved.   In the particular illustration used in this section, additional problems are involved, since the data used are a time series.   These problems, however, are not peculiar to the graphic method, and will be explained in the final chapter of this book.

Selected References

L. H. Bean: "A Simplified Method of Graphic Curvilinear Correlation," *Journal of the American Statistical Association*, Volume XXIV, December 1929, pp. 386–397. "Application of a Simplified Method of Correlation to Problems in Acreage and Yield Variations," *Journal of the American Statistical Association*, Volume XXV, December 1930, pp. 428–439.

L. H. Bean and G. B. Thorne: "The Use of Trend Residuals in Constructing Demand Curves," *Journal of the American Statistical Association*, Volume XXVII, March 1932, pp. 61–67. Graphical non-linear multiple correlation.

R. W. Burgess. *Introduction to the Mathematics of Statistics*, Chapter XIV; Houghton Mifflin Co., Boston, 1927.

B. H. Camp: *The Mathematical Part of Elementary Statistics*, Part II, Chapter VI; D. C. Heath and Co., Boston, 1934.

G. R. Davies and W. F. Crowder: *Methods of Statistical Analysis in the Social Sciences*, pages 264–280; John Wiley and Sons, New York, 1933.   Includes a graphic approximation method for multiple correlation.

H. T. Davis and W. F. C. Nelson: *Elements of Statistics*, Chapter XI; Principia Press, Bloomington, Indiana, 1935.   The approach is that of starting with the zero order coefficients and deriving the other measures from these.

Mordecai Ezekiel: *Methods of Correlation Analysis*, Chapters X–XXIII; John Wiley and Sons, New York, 1930.   These chapters include a detailed exposition of multiple correlation, both mathematical and graphic.

R. A. Fisher: *Statistical Methods for Research Workers* (Seventh Edition), pages 191–197; Oliver and Boyd, Edinburgh, 1936.   Deals with the logic of partial correlation.

F. C. Mills· *Statistical Methods Applied to Economics and Business* (Revised Edition), Chapter XVI; Henry Holt and Co., New York, 1938.

Henry Schultz: *The Theory and Measurement of Demand;* University of Chicago Press, Chicago, 1938.   An excellent advanced treatment of the application of correlation methods to a special branch of economics.   Part I is theoretical, while in Part II are given illustrations.   The reader should not attempt this book until he has read Chapter XXV of the present text, nor unless he has an acquaintance with the methods of calculus.

G. W. Snedecor: *Statistical Methods Applied to Experiments in Agriculture and Biology*, Chapter 13, Collegiate Press, Ames. Iowa, 1937.   Includes an explanation of multiple co-variance in groups.

L. H. C. Tippett: *The Methods of Statistics* (Second Edition), Chapter XI; Williams and Norgate, Ltd., London, 1937.

A. E. Waugh: *Elements of Statistical Method*, Chapter XI; McGraw-Hill Book Co.. New York, 1938.   A simple discussion of multiple correlation, linear and non-linear, and of joint correlation.   Partial correlation is not discussed.

G. U. Yule and M. G. Kendall: *An Introduction to the Theory of Statistics* (Eleventh Edition), Chapter 14; Charles Griffin and Co., London, 1937.

CHAPTER XXV

CORRELATION OF TIME SERIES AND FORECASTING

Correlation of Time Series

Preliminary adjustment of data. The technique ordinarily employed in the correlation of time series is exactly the same as that of correlating any other kind of data, the sole difference being in the nature of the variations compared. Let us take as an illustration the production and price of tame hay in the United States, by years, 1900–1936. The data are shown in the first three columns of Table 177, and are plotted as time series by solid lines in Chart 244, and as a scatter diagram in Chart 245. It is apparent from the latter chart that the correlation between the two series is negligible. Yet the two sections of Chart 244 indicate that there is rather high *negative* correlation between the short term movements of these series. On the other hand, the trends of the two series (shown by broken lines) are *positively* correlated. The result is that the positive long term relationship almost exactly cancels the negative short term relationship.

Since the trends can be compared by comparing their trend equations, it seems more fruitful to correlate, not the total movements of the two series, but only their short term movements. Sometimes this is done by correlating the year-to-year changes (first differences). These annual changes are shown in Table 177, and in the scatter diagram Chart 246. The coefficient of correlation is −.64. It is obvious that a decrease in price is associated with an increase in production, while an increase in price is associated with a decrease in production. This is as we should expect. Nevertheless, there are logical objections to the correlation of absolute changes. First, relating each value to the preceding year only partially adjusts for trend in price or quantity. Second, an increase in production from an abnormally low point would have a different effect on price than an increase of the same magnitude from a level which was already above normal. Finally, an increase in production from a low level, regardless of whether it was above or below normal, would affect price differently than

would a similar absolute increase from an absolutely high level. This third logical difficulty can be overcome by correlating percentages of preceding year rather than first differences. Percentages of preceding year are cal-

TABLE 177

PRODUCTION AND PRICE OF TAME HAY, ABSOLUTE CHANGE, AND PER CENT OF PRECEDING YEAR, BY YEARS 1900–1936

Year	Production (millions of tons)	Price per ton (dollars)	Change from preceding year		Per cent of preceding year	
			Production	Price	Production	Price
1900	49.8	9.78			107	101
1901	53.1	9.88	3 3	0.10	111	92
1902	59.1	9.05	6.0	−0.83	108	101
1903	63.6	9 18	4.5	0 13	103	96
1904	65.6	8 82	2.0	−0 36	102	96
1905	66.6	8 49	1.0	−0.33	91	122
1906	60.4	10 40	− 6.2	1.91	110	112
1907	66.3	11.60	5 9	1.20	108	78
1908	71.6	9.08	5.3	−2.52	96	116
1909	68.8	10 50	− 2.8	1.42	91	116
1910	62.9	12.16	− 5.9	1.66	83	119
1911	52.1	14.41	−10 8	2.25	133	81
1912	69.1	11.68	17.0	−2.73	90	106
1913	62.3	12.36	− 6.8	0.68	106	90
1914	65.8	11.11	3.5	−1.25	111	96
1915	73.3	10.65	7.5	−0.46	111	105
1916	81.2	11.18	7.9	0.53	88	153
1917	71 1	17.08	−10.1	5.90	96	118
1918	68.5	20.07	− 2.6	2.99	112	100
1919	76 6	20.15	8.1	0 08	99	88
1920	76.2	17.78	− 0.4	−2.37	93	68
1921	71.0	12.09	− 5 2	−5.69	114	104
1922	80.8	12.55	9.8	0.46	93	112
1923	75.3	14.10	− 5.5	1.55	105	98
1924	78.9	13.80	3.6	−0.30	85	101
1925	67.3	13.95	−11.6	0.15	100	101
1926	67.1	14.08	− 0.2	0.13	124	80
1927	83.3	11.30	16.2	−2.78	87	108
1928	72.2	12.22	−11.1	0.92	105	100
1929	76.1	12.19	3.9	−0.03	84	104
1930	64.0	12.62	−12.1	0.43	104	72
1931	66.6	9.03	2.6	−3.59	108	74
1932	71.8	6.65	5.2	−2.38	93	122
1933	66.5	8.11	− 5.3	1.46	83	172
1934	55.3	13.95	−11.2	5.84	141	56
1935	78.1	7.80	22.8	−6.15	81	146
1936	63.3	11.39	−14.8	3.59		

Source Production, 1900–1923 United States Department of Agriculture, Bureau of Agricultural Economics, *Revised Estimates of Tame Hay Acreage, Yield and Production, 1896–1929* (mimeograph bulletin) p 3, 1924–1936 United States Department of Agriculture, *Agricultural Statistics, 1937,* p 224 Price, 1900–1918 Information provided by Bureau of Agricultural Economics, 1919–1931 United States Department of Agriculture. *Yearbook of Agriculture, 1935,* p 534, 1935–1936 *Agricultural Statistics 1937,* p. 226

culated in the last two columns of Table 177, and Chart 247 is a scatter diagram of the results. The correlation coefficient is −.66. The picture is not greatly different, except that the relationship now is perhaps non-linear.

In order to make a more logical comparison, we must correlate, not percentages of the preceding year, but percentages of normal. In the present instance it seems best to make still another adjustment in the price series. Besides the price trend that is characteristic of hay in particular, we have changes in the series which are associated with changes
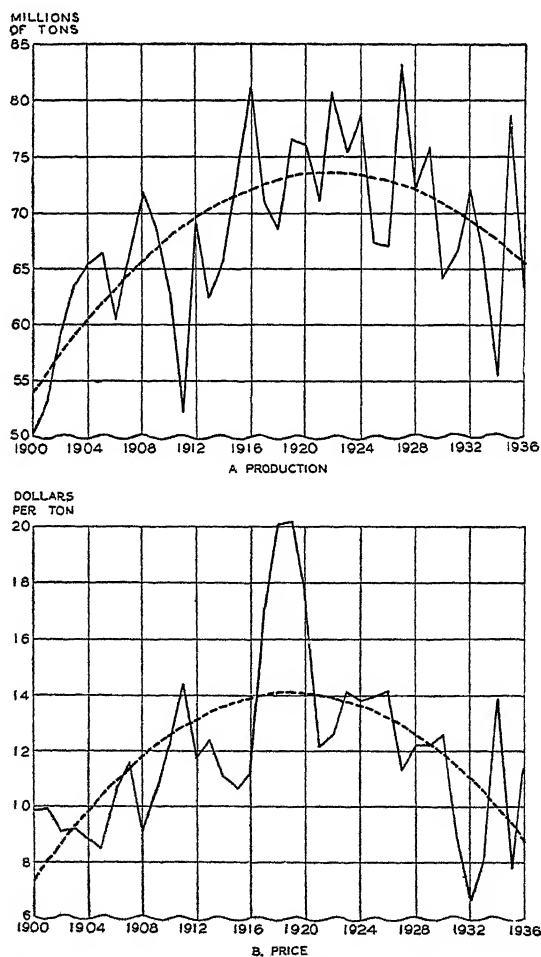
Chart 244. Production and Price of Tame Hay in the United States, and Secular Trends, by Years, 1900-1936. (Data of Table 177.)

in commodity prices in general. To adjust for both of these factors, we first divide the price of hay by an index of commodity prices. Although it is difficult to say what index is most appropriate, the United States Bureau of Labor Statistics Index of Wholesale Prices has been used. The adjusted data may now be referred to as expressing changes in the *real price* of tame hay. The second step is to compute a trend for the real price series. The progress to this point is shown in Chart 248. The trend is a third degree curve. The final step is to divide the real price series by

PRICE
DOLLARS PER TON
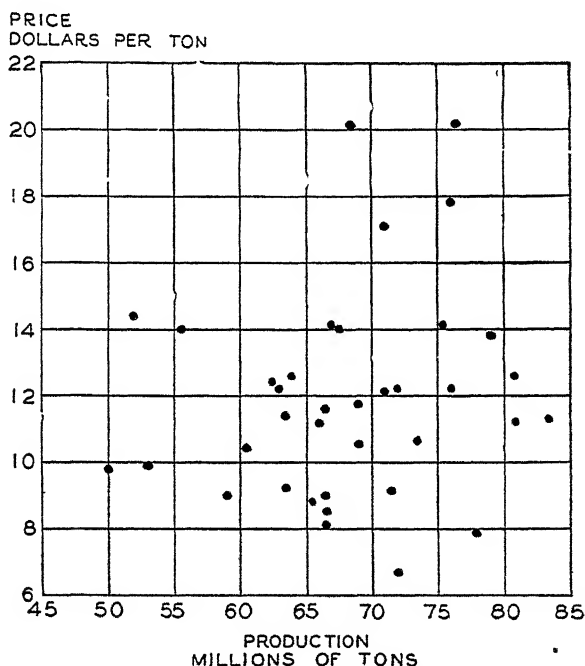
Chart 245. Scatter Diagram of Production and Price of Tame Hay, by Years, 1900–1936. (Data of Table 177.)

the trend values. These three steps are shown in Table 178, columns 5–9. It would be entirely logical to adjust hay production for changes in the number of animal units, and not entirely unreasonable to adjust the hay production figures for changes in general production in the United States, or in production of commodities which are substitutes for hay.

For purposes of this illustration, however, no such adjustment has been made. A second degree curve has been fitted directly to the data, and they have then been divided by the trend values. The numerical results are given in Table 178, columns 2–4. A comparison of the two adjusted time series is afforded by Chart 249, and by the scatter diagram Chart

250. The correlation coefficient is $r = -.79$, which is materially higher than that obtained by the other two methods.

Another method of dealing with the trend factor when correlating time

TABLE 178

ADJUSTMENT OF PRODUCTION AND PRICE OF TAME HAY DATA FOR PURPOSES OF CORRELATION, 1900–1936

| Year | Production | | | Price | | | | |
| | X | X_c | Per cent of trend $[X \div X_c]$ | Nominal price | Price index [1926 = 100] | Real price Y [Col 5 $\div$ Col 6] | Y_c | Per cent of trend $[Y \div Y_c]$ |
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1900	49 8	53.8	92	9.78	56.1	17.43	15.78	110
1901	53.1	55.6	96	9.88	55.3	17.87	16.09	111
1902	59.1	57.3	103	9.05	58.9	15.37	16 33	94
1903	63 6	58 8	108	9.18	59.6	15.40	16.52	93
1904	65.6	60 4	109	8.82	59.7	14.77	16.65	89
1905	66.6	61 8	108	8.49	60.1	14.13	16.73	84
1906	60.4	63.1	96	10.40	61 8	16.83	16 76	100
1907	66.3	64 4	103	11.60	65.2	17.79	16.75	106
1908	71.6	65 6	109	9.08	62.9	14.44	16.63	87
1909	68.8	66.6	103	10.50	67.6	15.53	16.60	94
1910	62.9	67.6	93	12.16	70.4	17.27	16.48	105
1911	52.1	68 6	76	14.41	64.9	22.20	16.32	136
1912	69.1	69.4	100	11.68	69.1	16.90	16.14	105
1913	62.3	70.2	89	12.36	69.8	17.71	15.94	111
1914	65.8	70.9	93	11.11	68.1	16.31	15.71	104
1915	73.3	71.5	102	10.65	69.5	15.32	15.48	99
1916	81.2	72.0	113	11.18	85.5	13.08	15.23	86
1917	71.1	72 4	98	17.08	117.5	14.54	14.97	97
1918	68.5	72.8	94	20.07	131.3	15.29	14.72	104
1919	76.6	73.1	105	20.15	138.6	14.54	14.46	101
1920	76.2	73.3	104	17.78	154.4	11.52	14.20	81
1921	71.0	73.4	97	12.09	97.6	12.39	13.95	89
1922	80.8	,73.4	110	12.55	96.7	12.98	13.72	95
1923	75.3	73.4	103	14.10	100.6	14.02	13.50	104
1924	78.9	73.2	108	13.80	98.1	14.07	13.29	106
1925	67.3	73.0	92	13.95	103.5	13.48	13.11	103
1926	67.1	72.7	92	14.08	100.0	14.08	12.96	109
1927	83.3	72.4	115	11.30	95.4	11.84	12.83	92
1928	72.2	71.9	100	12.22	96.7	12.64	12.74	99
1929	76.1	71 4	107	12.19	95.3	12.79	12.69	101
1930	64.0	70.7	91	12.62	86.4	14.61	12.68	115
1931	66.6	70.0	95	9.03	73.0	12.37	12.71	97
1932	71.8	69.2	104	6.65	64.8	10.26	12.79	80
1933	66.5	68.4	97	8.11	65.9	12.31	12.92	95
1934	55.3	67.4	82	13.95	74.9	18.62	13.11	142
1935	78.1	66.4	118	7.80	80.0	9.75	13.36	73
1936	63.3	65.3	97	11.39	80.8	14.10	13.67	103

Source: See Table 177   Price Index is United States Bureau of Labor Statistics Index of Wholesale Prices, published in United States Bureau of Labor Statistics, *Wholesale Prices 1931*, p 14, and *Wholesale Prices, December and Year 1937*, p 3

series is to introduce a third variable, time, and employ multiple correla-
tion analysis. This is done instead of eliminating the trend, and has the
added advantage (when dealing with problems like the one under discus-
sion) of showing the net annual change in price which is a result of changing
demand.

The above illustration employed annual data. Had monthly data been
used, it would have been desirable to have deseasonalized the data also,
since the presence of violent seasonal movements might have distorted
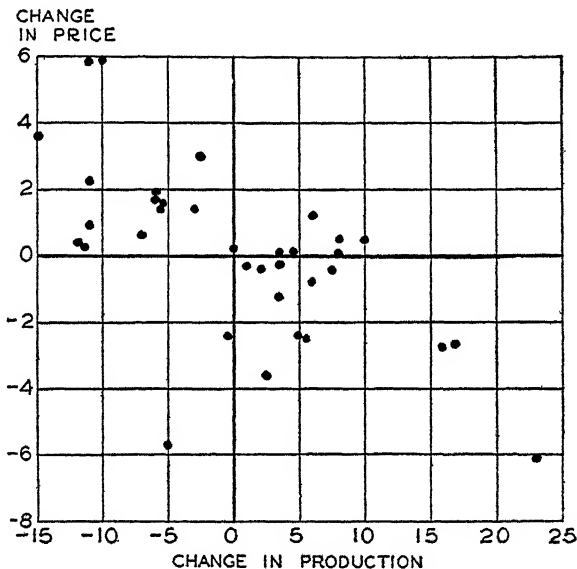
Chart 246. Scatter Diagram of Change from Preceding Year of Production and
Price of Tame Hay, 1900-1936. (Although there are 36 observations, only 35 dots are
visible since two observations coincide. Data of Table 177.)

the correlation of the cyclical or other short term movements being com-
pared. This procedure will be treated later in this chapter.

Correlation of adjusted cyclical relatives. A comparison of the cycles
of production and real price of tame hay expressed as percentages of normal,
as shown in Chart 249, may readily be made. However, the graphic com-
parison of two series that differ greatly in amplitude is difficult. Thus,
although we can tell by inspection of section A of Chart 251 that the turn-
ing points of passenger car production and electric power production are
the same, the two curves are at times so far apart that it is difficult to
judge how closely they are associated throughout. Mathematically, of
course, the closeness of relationship may be ascertained by computing the

coefficient of correlation by the customary product-moment formula, as in Table 179, part A.

For graphic visualization, however, it is helpful to make two further adjustments in the data before plotting. These adjustments, which are embodied in section B of Chart 251, are as follows. First, the data are converted into percentage deviations from normal by subtracting 100 (or the actual mean) from each per cent of normal. If the trend has been fitted to the data by the method of least squares for a period exactly coin-

PRICE
PER CENT OF
PRECEDING YEAR
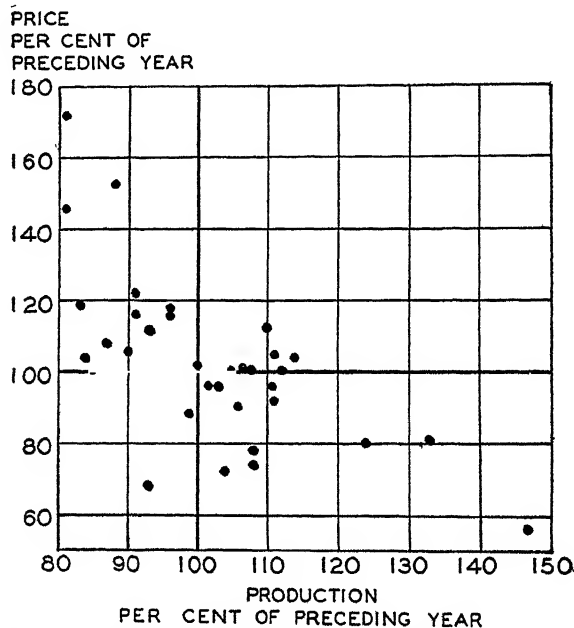
Chart 247. Scatter Diagram of Per Cent of Preceding Year of Production and Price of Tame Hay, 1900-1936. (Data of Table 177.)

ciding with the period under comparison, the mean of the series will be approximately 100 per cent. If the trend covers a longer period than do the data being correlated, or if the trend has been extended, or if some method other than least squares has been used (as in the present instance), it will generally be advisable to use deviations from the actual mean rather than from 100 per cent (in order that $\Sigma x = 0$ and $\Sigma y = 0$), if the method of correlation about to be explained is to be used. In the present instance deviations are taken from the mean. Second, each series is expressed in units of its standard deviation. (Sometimes the average deviation is used instead.) Thus the standard deviation of electric power production has been computed as in Table 179, part B, and the deviation for each year
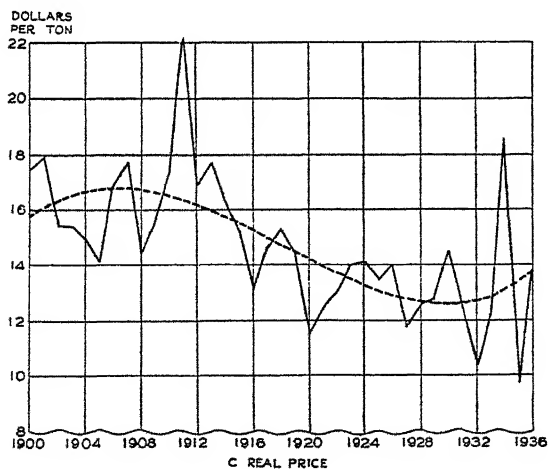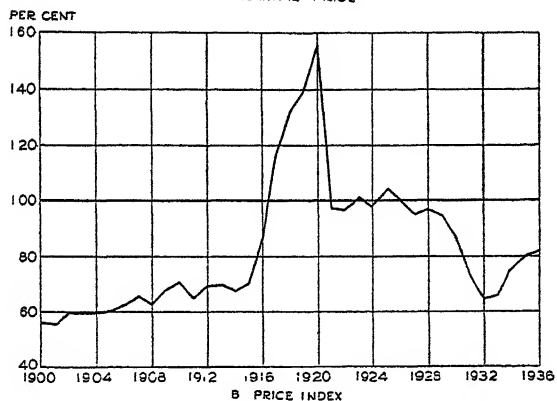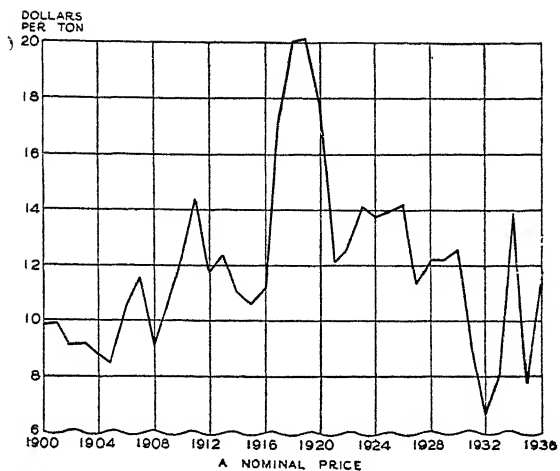
Chart 248. Nominal Price of Tame Hay, Index of Wholesale Commodity Prices, Real Price, and Trend of Real Price, by Years, 1900-1936. (Data of Table 178.)

TABLE 179

A. CORRELATION OF CYCLICAL MOVEMENTS OF ELECTRIC POWER PRODUCTION AND PASSENGER CAR PRODUCTION, 1921-1932, BY PRODUCT-MOMENT METHOD

Year (1)	Electric power production X (2)	Passenger car production Y (3)	XY (4)	X² (5)	Y² (6)
1921	92.55	73.56	6,807.98	8,565.50	5,411.07
1922	96.62	89.64	8,661.02	9,335.42	8,035.33
1923	102.45	116.37	11,922.11	10,496.00	13,541.98
1924	97.77	92.83	9,075.99	9,558.97	8,617.41
1925	98.44	109.30	10,759.49	9,690.43	11,946.49
1926	100.44	111.26	11,174.95	10,088.19	12,378.79
1927	99.11	84.03	8,328.21	9,822.79	7,061.04
1928	103.17	110.65	11,415.76	10,644.05	12,243.42
1929	108.93	134.85	14,689.21	11,865.74	18,184.52
1930	105.27	87.58	9,219.55	11,081.77	7,670.26
1931	98.72	66.72	6,586.60	9,745.64	4,451.56
1932	87.82	41.49	3,643.65	7,712.35	1,721.42
Total	1,191.29	1,118.28	112,284.52	118,606.85	111,263.29

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}}$$

$$= \frac{12(112,284.52) - (1,191.29)(1,118.28)}{\sqrt{[12(118,606.85) - (1,191.29)^2][12(111,263.29) - (1,118.28)^2]}}$$

$$= +.816.$$

TABLE 179 (Continued)

B. CORRELATION OF CYCLICAL MOVEMENTS ADJUSTED FOR AMPLITUDE OF ELECTRIC POWER PRODUCTION AND PASSENGER CAR PRODUCTION, 1921–1932, BY PRODUCT-MOMENT METHOD

Year	Per cent deviation* x	x^2	$\frac{x}{\sigma_x}$	Per cent deviation† y	y^2	$\frac{y}{\sigma_y}$	$\frac{x}{\sigma_x} \times \frac{y}{\sigma_y}$
1921	− 6.724	45.212	−1.258	−19.63	385.34	− .810	1.019
1922	− 2.654	7.044	− .497	− 3.55	12.60	− .146	.073
1923	− 3.176	10.087	.594	23.18	537.31	.956	.568
1924	− 1.504	2.262	− .281	− .36	.13	− .015	.004
1925	− .834	.696	− .156	16.11	259.53	.665	− .104
1926	1.166	1.360	.218	18.07	326.52	.745	.162
1927	− .164	.027	.031	− 9.16	83.91	− .378	.011
1928	3.896	15.179	.729	17.46	304.85	.720	.525
1929	9.656	93.238	1.807	41.66	1,735.56	1.719	3.106
1930	5.996	35.952	1.122	− 5.61	31.47	− .231	− .259
1931	− .554	.307	− .104	−26.47	700.66	−1.092	.114
1932	−11.454	131.194	−2.144	−51.70	2,672.89	−2.133	4.573
Total. .	…	342.558	…	…	7,050.77	…	+9.792
Mean. .	…	28.547	…	…	587.56	…	+ .816
σ	…	5.343	…	…	24.24	…	…

* X values of section A, column 2 = 99.274. $\bar{X} = 1,191.29 \div 12 = 99.274$.
† Y values of section A, column 3 = 93.190. $\bar{Y} = 1,118.28 \div 12 = 93.190$.

$$r = \frac{1}{N}\Sigma\left(\frac{x}{\sigma_x} \times \frac{y}{\sigma_y}\right) = \frac{+9.792}{12} = + .816$$

Source: United States Department of Commerce, Survey of Current Business, 1936 Supplement and subsequent issues. The trend fitted to electric power production is not that computed in Chapter XV but is a high-low mid-point trend. This type of trend was used also for passenger car production, and is the same as that given in Table 92.

has been divided by this value (5.343).   A similar procedure has been fol-
lowed for passenger car production; the standard deviation for this series
is much larger (24.24).   When each series has been divided by its own
standard deviation, the two resulting series show the same degree of fluc-
tuation.   It is now much easier to compare the two series graphically.
The degree of conformance is seen in part B of Chart 251 to be high.   Now,
r is most easily computed by the expression

$$r = \frac{1}{N} \Sigma\left(\frac{x}{\sigma_x} \times \frac{y}{\sigma_y}\right).$$

This formula was given on page 666, note 13.   (If the data have been con-
verted into average deviation units instead of standard deviation units
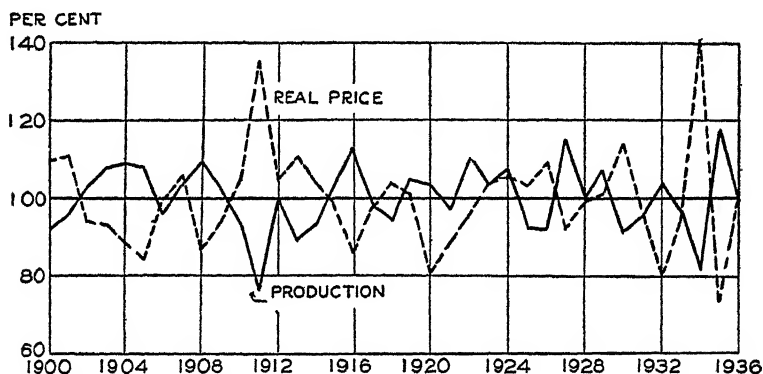
Chart 249.   Production and Real Price of Tame Hay, as Percentages of Trend, 1900–
1936.   (Data of Table 178.)

the formula on p. 804, note 1, should be used.)   Making our substitution
from part B of Table 179, we find

$$r = \tfrac{1}{12}\,(+9.792) = +.816.$$

Any of the formulae for r with which the reader is now familiar could be
used instead of this one, but the above is easiest if each series has already
been adjusted for amplitude by dividing through by its standard devia-
tion for purposes of graphic comparison.   Of course, it will not always, or
even usually, be found advisable to express the data in deviation form
and to adjust the series for differences in amplitude by dividing each
series by its respective standard deviation.   It is much more laborious to
make the adjustments and then correlate, than it is to correlate by the
customary method shown in part A of Table 179.   Incidentally, this table
shows that the same value of r is obtained when correlating the data in
terms of σ, as when per cent of trend values are correlated.

    An interesting by-product is available when this method of correlation

is used.   It was explained on page 666, note 13, that r could be thought of as the slope of the estimating line b_{yx} when each series has been expressed in terms of its own standard deviation, that is, $b\sigma_y\sigma_x$.   This procedure has literally been followed in the present instance, and in Chart 252

REAL PRICE
PER CENT OF NORMAL
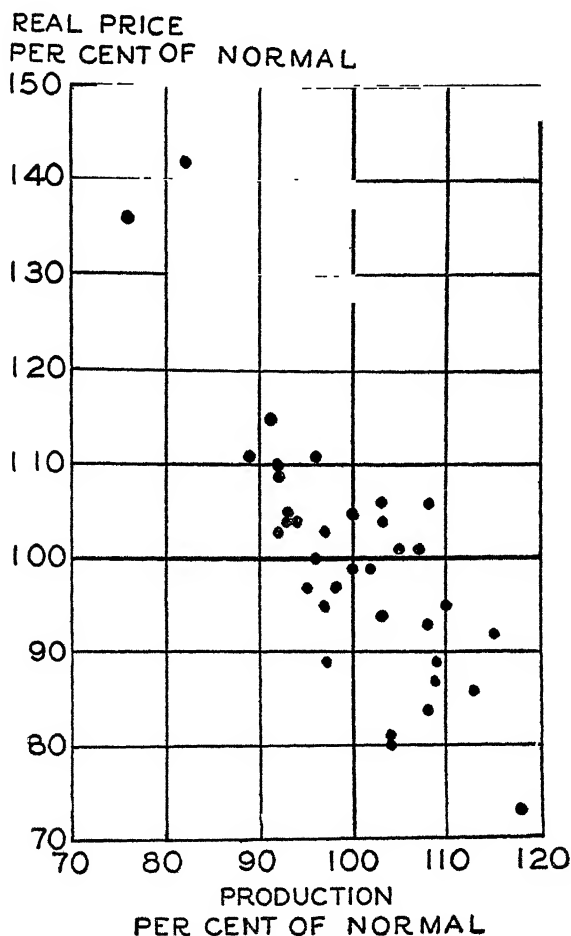
Chart 250.   Scatter Diagram of Production and Real Price of Tame Hay as Percentages of Trend, 1900-1936.   (Although there are 37 observations, only 36 dots are visible since two observations coincide.   Data of Table 178.)

the adjusted data have been plotted and the line $\dfrac{y_c}{\sigma_y} = .816\,\dfrac{x}{\sigma_x}$ has been shown.   This chart gives visual evidence of the correctness of the above way of looking at r.

PER CENT

PASSENGER CAR

ELECTRIC POWER

A

STANDARD
DEVIATIONS

PASSENGER CAR

ELECTRIC POWER

B
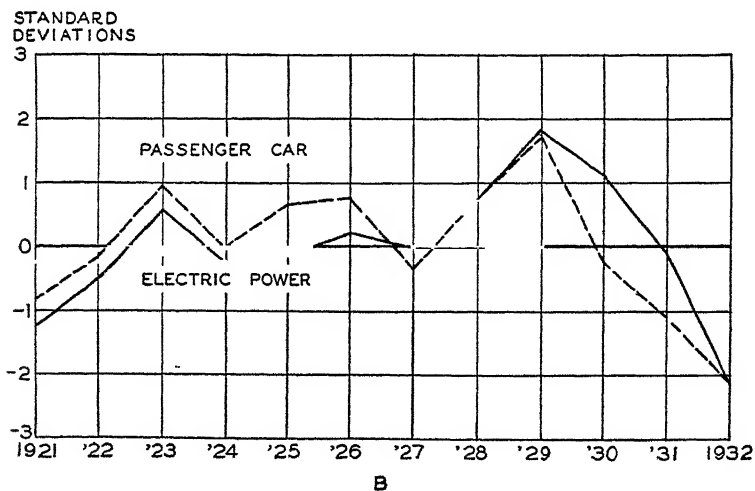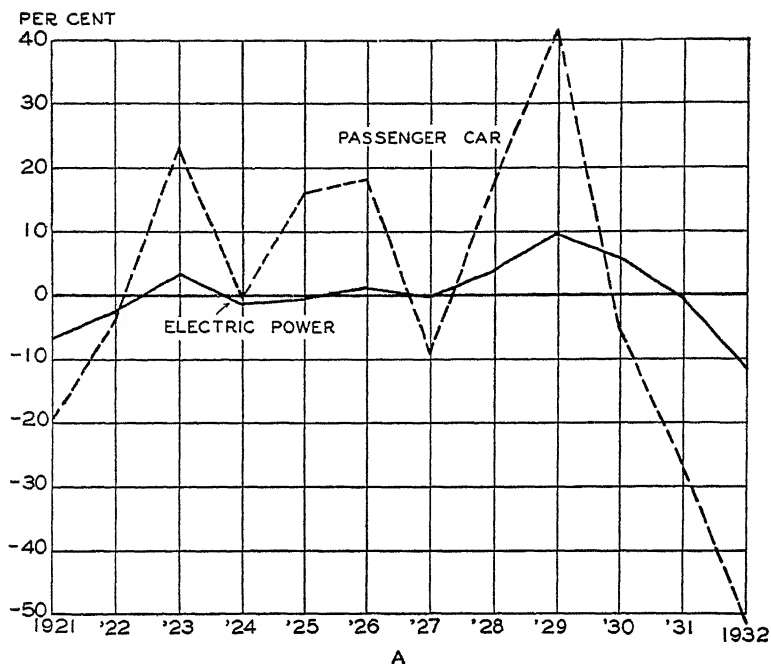
Chart 251. Passenger Car Production and Electric Power Production (A) as Percentage Deviations from Trend and (B) as Deviations from Trend in Units of Their Standard Deviations, 1921-1932. (Data of Table 179.)

Problems in correlating time series. It must be evident that the value of the correlation coefficient is affected by the type of trend fitted to the data, and the period to which it is fitted. If a period of 10 years is being correlated, it would not be logical to use for one series a section of a trend fitted over a 100-year period and for the other a trend fitted to data extending over 10 years only. The former trend would, in all likelihood,

PASSENGER CAR
PRODUCTION
UNITS: σ_y

$b_{\sigma_y \sigma_x} = +.816 \longrightarrow$

ELECTRIC POWER PRODUCTION
UNITS: σ_x
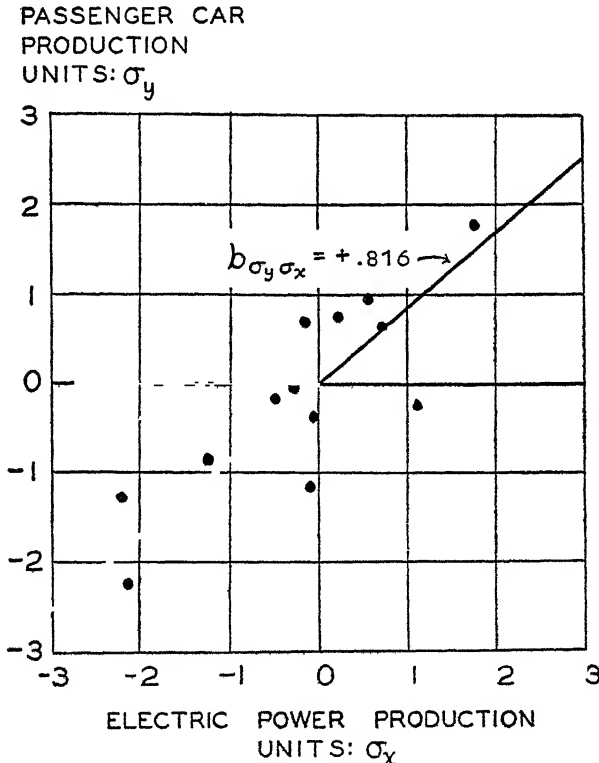
Chart 252. Scatter Diagram of Passenger Car Production and Electric Power Production as Percentage Deviations from Trend in Units of Their Standard Deviations. (Data of Table 179B.)

fail to pass through the approximate center of each cycle, and might not even touch some of the cycles. Consequently the correlation coefficient might understate or overstate the degree of relationship between the series. It must also be apparent that the use of an inflexible trend for one series and a flexible trend for the other would produce similar results. If we wish to correlate cyclical movements, it seems best therefore to use a trend that goes approximately through the center of each cycle. It may

be that no simple mathematical curve will be satisfactory and that some relatively subjective method, such as the high-low mid-point method, will have to be resorted to, at least as a first approximation.

Another problem to consider is whether the Pearsonian method of correlation, based on the second moments, is appropriate for correlating time series. The fluctuations of a time series are not usually distributed normally around the trend line. There are often a few extreme deviations, which, when squared, largely determine the value of r. In our last illustration, the distribution is fairly regular, although 1929 is rather high and 1932 is exceptionally low. With this problem in mind, some authorities suggest the use of the rank method when the extreme deviations are particularly large. Another solution is the use of a formula based on first moments, rather than second.[1]   In view of the fact that interest frequently centers in whether two series are moving in the same general direction (positive or negative) at the same time, without regard to the magnitude either of their level or their change, it may be that none of the orthodox methods of correlation are satisfactory.

A further difficulty in correlating time series is that we have no logical basis for estimating the reliability of the coefficient of correlation. The chief objection to the use of any reliability test for r for time series is that the different observations are not randomly distributed—each observation in a time series is related to values in that series for preceding and subsequent points of time. Furthermore, we cannot generalize concerning the exact nature of this interrelationship, and hence we cannot develop any general theory of reliability applicable to this branch of statistics. Perhaps this difficulty will become more obvious when we ask how many

---

[1] See "The Validity of Correlation in Time Sequences and a New Coefficient of Similarity," by O. Gressens and E. D. Mouzon, Jr., *Journal of the American Statistical Association*, Vol. XXII, December 1927, pp. 483-492. This method is further elucidated and its relation to r explained by George R. Davies, in an article entitled "First Moment Correlation," appearing in the *Journal of the American Statistical Association*, Vol. XXV, December 1930, pp. 413-427. The formula is

$$C_2 = \frac{\Sigma s(2N - \Sigma|s|)}{N^2},$$

where s refers to the smaller of each pair of items when each series is expressed as deviations from the mean in terms of average deviations $\left(\dfrac{x}{\mathrm{AD}_x} \text{ and } \dfrac{y}{\mathrm{AD}_y}\right)$. When summing algebraically, s is positive if the signs of the paired deviations are alike, and negative if they are unlike. Using the x and y data of Table 179, part B, we find that

$$C_2 = \frac{7.70(24 - 8.68)}{144} = +.819.$$

The computational labor is much less than that involved in using the formula on p. 800 Davies also explains certain short cuts in computation.

independent observations are contained in the cyclical relatives used in the last illustration. Although there are 12 years, there are not 12 degrees of freedom. There are only three complete cycles (measuring from trough to trough). Are there then only 3 independent observations? (Subtracting 2 more for the constants a and b in the estimating line leaves only 1 degree of freedom.) But there are more than 3, since each observation in a cycle is not completely dependent on the preceding values. If we now had monthly data, would we have 144 independent observations for the 12 years? Of course not. But, how many we would have it is impossible to say.

Measurement of Lag

It is an aid to the understanding of economic processes to measure the period of time by which one series precedes another. For a business man it is especially profitable to be able to predict, a number of months in advance, when business will pick up or recede. Since the turning points in all time series do not occur at the same time, it is necessary to pick out a series which precedes, with some degree of regularity in its turning points, the series we wish to predict, and observe how many months' interval there is between the turning points of the two series. In order to do this precisely, the device of correlation is frequently resorted to. The two series having been reduced to per cent of normal, they are plotted on separate sheets of graph paper, with scales so chosen that the amplitudes of fluctuation will be about the same. These sheets are then placed together and held up to a light, and when they are slid back and forth, some point is reached at which the correspondence is closest.

In the accompanying illustration, the Index of Industrial Production of the Board of Governors of the Federal Reserve System is to be forecast by an index of production of durable consumers goods constructed by the writers. Both series have been adjusted for trend and seasonal variation, and the former has been slightly smoothed. Chart 253 shows the two series superimposed: (A) with no lag of either series; and (B) with the series to be forecast moved two months to the left, so that January 1919 of the durable consumers goods series is even with March 1919 of the Federal Reserve index. This last position seems to show the best correspondence between the two series; it is the best visual estimate of the lag. In other words, increases in durable consumers goods production seem to precede increases in general industrial production by about two months. With the 2-month lag of industrial production, $r = +.933$. Computations are shown in Table 180.

It may be, however, that better correlation would have been obtained with some other period of lag. Logically it may even be hypothecated

that production of durable goods for consumers waits on purchasing power derived from a general industrial pickup.   The various values for r with different lag assumptions are as follows:

If production of durable consumers goods lags behind industrial production:

		r
10 months	.	+.729
9 months	. . . .	+.743
8 months	.	+.764
7 months	.	+.778
6 months		+.794
5 months	.	+.818
4 months		+.833
3 months	. .	+.857
2 months		+.887
1 month	. . . .	+.908
No lag	. .	+.924

If industrial production lags behind production of durable consumers goods:

		r
1 month.	. . . .	+.932
2 months...	.. . .	+.933
3 months	. . . .	+.929
4 months	. . . .	+.923
5 months	. . . .	+.917
6 months ..	... .	+.902
7 months . .	. . . . . .	+.899
8 months .	. . . .	+.893
9 months	. . ..	+.870
10 months	. . . .	+.850
11 months .	. . ·..	+.820
12 months	... .	+.785
13 months	....	+.753
14 months.	.. . .	+.719

The results are shown graphically in Chart 254.   It may be concluded that production of durable consumers goods forecasts changes in industrial production by *about* two months, and that industrial production is a less satisfactory forecaster of durable consumers goods production.

It would be useless to compute σ_r for these data, and worse than useless to make the Z transformation and attempt to evaluate the significance of the difference between these r values.   As has already been explained, the interdependence of the different time series observations invalidates the usual procedures for judging the reliability of the correlation coefficient. Nor is it clear how many degrees of freedom are sacrificed when such a lag adjustment is made.[2]   Just as a random sample from an uncorrelated

---

[2] The authors obtained a correlation of $-.84$ from random data of 27 items adjusted for a 3-period lag   See Croxton and Cowden, *Practical Business Statistics* pp. 457–460, Prentice-Hall, Inc., New York, 1934.

population is likely to show some correlation, so any sample which is uncorrelated when taken synchronously will show correlation when adjusted for lag. It is better, therefore, not to attempt a mathematical estimate of the reliability of r. It is advisable, however, to include a considerable period of time in the series, and to compute r for various
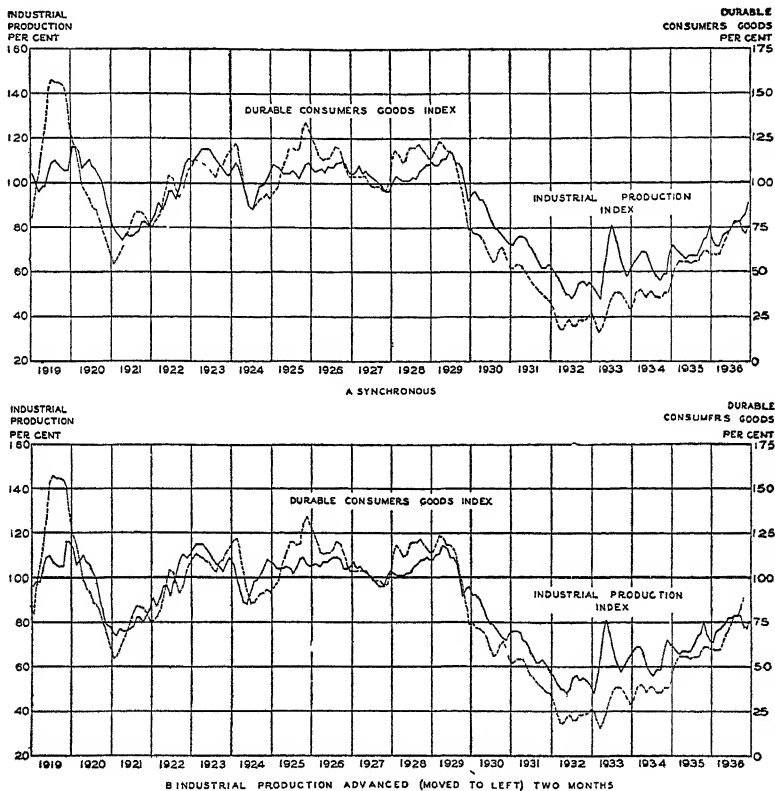
Chart 253. Cyclical Movements of Durable Consumers Goods Production Index and Federal Reserve Index of Industrial Production, 1919-1936: A. Synchronous, B. With Industrial Production Moved 2 Months to the Left. (The time scale refers to durable consumers goods. Data of Table 180.)

sub-periods of time before coming to anything but the most tentative conclusion. However, even if this procedure gives the statistician confidence that he has discovered a real relationship, it does not necessarily follow that the equation will be useful in forecasting the future. Each trend, deviations from which are correlated, must be capable of being extrapolated without serious error. Also, economic institutions and condi

tions are constantly changing, and what may have been an important relationship in the past may be either of a different nature or of smaller relative importance in the future.

The equation for estimating industrial production, Y, from durable consumers goods production, X, may be obtained by the formula

$$(Y_C - \overline{Y}) = r \frac{\sigma_y}{\sigma_x} (X - \overline{X}).$$

TABLE 180

CORRELATION OF FEDERAL RESERVE INDEX OF INDUSTRIAL PRODUCTION AND DURABLE CONSUMERS GOODS PRODUCTION WITH 2-MONTH LAG OF INDUSTRIAL PRODUCTION, 1919–1936

Year and month (for durable consumers goods)*	Durable consumers goods production X	Industrial production Y	XY	X^2	Y^2
1919:					
January	80	96	7,680	6,400	9,216
February	97	98	9,506	9,409	9,604
March.	107	98	10,486	11,449	9,604
April	119	104	12,376	14,161	10,816
May.	133	109	14,497	17,689	11,881
June.	153	110	16,830	23,409	12,100
July	157	107	16,799	24,649	11,449
August	156	106	16,536	24,336	11,236
September	156	105	16,380	24,336	11,025
October	154	105	16,170	23,716	11,025
November	149	116	17,284	22,201	13,456
December	135	116	15,660	18,225	13,456
1936:					
January	60	71	4,260	3,600	5,041
February	60	76	4,560	3,600	5,776
March	60	77	4,620	3,600	5,929
April	66	79	5,214	4,356	6,241
May	70	82	5,740	4,900	6,724
June	74	82	6,068	5,476	6,724
July	78	83	6,474	6,084	6,889
August	78	83	6,474	6,084	6,889
September	78	86	6,708	6,084	7,396
October	73	91	6,643	5,329	8,281
Total	17,612	18,984	1,692,884	1,703,842	1,761,028

* Durable consumers goods production (X) is from January 1919 through October 1936   Industrial production (Y) is from March 1919 through December 1936   See p 805 for explanation
Source: Durable Consumers Goods Production Index was computed by the authors   The Federal Reserve Industrial Production Index (Manufacturing and Minerals) was taken from *Standard Trade and Securities, Statistics*, Vol 3, p D-4; and *Current Statistics*, May 1937

Substituting in this equation gives the equation

$$Y_C = 46.485 + .51307X.$$

This equation may be used for making forecasts of cyclical movements for individual months. Thus, since the index number for durable consumers goods production stood at 80 for February 1937, our best estimate for industrial production for April 1937 is

$$Y_C = 46.485 + .51307(80) = 87.53.$$
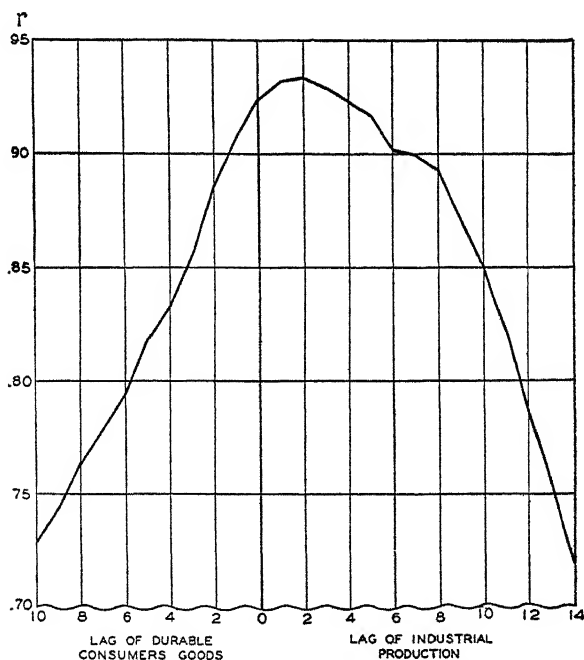
Chart 254. Values of r for Durable Consumers Goods Production and Industrial Production with Different Assumptions of Lag. (For data see p 806.)

The actual index number for April 1937 was 88. Such a close agreement is unusual, however. The standard error of estimate

$$\sigma_{y_s} = \sigma_y\sqrt{1 - r^2} = 6.83.$$

This indicates that, if estimates had been made from the equation for each month included in the period covered by the data, about two-thirds of the estimates would have been in error by less than 6.83. It must be remembered, however, that we are not dealing with variables which follow a chance distribution, and therefore σ_{y_s} should be thought of as only a very rough estimate of the range within which 68 per cent of the actual Y values fall.

The slowness with which economic data are reported and the scarcity of time series on a basis shorter than a month are factors that impair the usefulness of this method. It is quite possible that weekly, daily, or hourly data might bring to light relationships which are known and utilized only by a few "insiders." The theorist argues that all economic processes are interrelated. It does not seem logical that the cause-and-effect relationships which supposedly surround us on every side must always take a month or more for their development. There must be many that work out in a few days, a few hours, or nearly instantaneously. If the market hears that a new industrial use has suddenly been announced for copper, it does not wait weeks or even hours to show its reaction in a price change. As data are made available upon a weekly, daily, or more frequent basis, it is conceivable that very useful lags and leads may be obtained.

The use of correlation procedure in forecasting is subject to all the objections previously raised to its use in correlating time series, and others as well. These objections are:

(1) Being based on the second moments, r gives undue influence to the occasional extreme deviations characteristic of time series. In fact, some statisticians insist that a person's visual impression of lag is a more satisfactory measure than is r.

(2) The lag may be different at recession than it is at revival. As was mentioned in Chapter XIX, the National Bureau of Economic Research computes average lag or lead at revival and at recession with respect to its reference cycle.

(3) Interest often centers mainly on turning points, while r gives equal importance to lags at all phases of the cycle. It may be profitable to be able to foretell merely when to expect a change in direction, even though the amount of change cannot be forecast.

(4) It is a laborious process to compute r for a large number of lag hypotheses.

(5) In addition to criticisms of the coefficient of correlation as a measure of relationship, one may also criticize the nature of the variations correlated, arguing that a person can more accurately predict the future with respect to the present than he can with respect to some normal, which is often difficult to estimate correctly.

Distribution of lag. A refinement of the usual lag measurement has been introduced by Irving Fisher. It is his contention that the business cycle is largely a "dance of the dollar." But although price changes are a dominating factor in changes in the volume of trade and employment, a given price change does not have its entire effect in any one month

Rather, the effect is distributed over a number of months, reaching a climax after a certain period of time and then dwindling in importance, after the fashion of the ordinates of a frequency distribution. The type of distribution which Fisher considers most logical is one which becomes normal if the logarithm of time is taken as the abscissa, the origin being the month whose price change is under consideration. Furthermore, it is not the price level that is considered the causal factor, but the rate of price change. Thus the percentage price change for June would be approximately

$$\frac{(\text{July Price} - \text{May Price}) \div 2}{\text{June Price}}.$$

The mechanics of determining the best constants for the logarithmic frequency distribution (the mode and the standard deviation) involve a great deal of labor, the major part of which can be saved by simplifying the hypothesis slightly. If the maximum effect of a given price change is assumed to be the month (or other unit of time) following its occurrence, and the decline in influence is assumed to be linear, the statistical problem consists in determining only for how long a period of time the given cause exercises any effect whatever   This simplification of the problem does not usually affect materially the accuracy of the results. Following is a brief description of a method devised by Fisher for shooting forward and cumulating the effect of price changes.

Determine the best fixed lag by the usual correlation method. Upon the hypothesis that the duration N of the influence of price change is three times the fixed lag period, compute an estimating index by methods which will be explained shortly. Correlate the estimating index with the actual data of the series being estimated. Repeat this process with N equal to four times the fixed lag period. Try any such third hypothesis as seems best, and continue until the highest correlation is obtained.

A procedure for the actual computing of the forecasting index for any given hypothesis of lag is quoted from an article by Irving Fisher.[3] In order to preserve consistency with our other symbols, X is taken as the causal factor (price change), rather than a as used by Fisher. Also, t refers

---

[3] See "Note on a Short-Cut Method for Calculating Distributed Lags" by Irving Fisher in *Extrait du Bulletin De L'institut International de Statistique*, XXIX: 3. In this article Fisher describes further short-cuts in computation, and refers the reader for further discussion of principles to Max Sasuly, *Trend Analysis of Statistics*, Brookings Institution, 1934, pp. 134–135, 145, 149, 201, and Chapter X. Other articles on this subject by Irving Fisher are "Our Unstable Dollar and the So-called Business Cycle," *Journal of The American Statistical Association*, June 1925, Vol. XX, pp. 179–202, and "Changes in the Wholesale Price Index in Relation to Factory Employment," *Journal of the American Statistical Association*, September 1936, Vol. 31, pp. 496–506. This includes a discussion by Morris A. Copeland and a rejoinder by Irving Fisher.

to a given point of time. Thus the first month is t_1, the second month t_2, and so on. Quoting from Fisher (with minor alterations in wording):

Let us suppose that the single "cause" X_1, at month t_1, has its effect distributed over succeeding months (t_2 t_3, etc) in diminishing degrees proportional to the numbers 6, 5, 4, 3, 2, 1, ceasing after the sixth month (that is, after t_7). While the total effect of X_1 is proportional to X_1 itself, the part of this total which is felt at time t_2 is

$$\frac{6}{6+5+4+3+2+1} \text{ of } X_1, \text{ or } \frac{6}{21} \text{ of } X_1;$$

and the parts in the five succeeding months are respectively

$$\frac{5}{21} \text{ of } X_1, \frac{4}{21} X_1, \frac{3}{21} X_1, \frac{2}{21} X_1, \frac{1}{21} X_1.$$

Similarly the "cause" X_2, at month t_2, produces its effect in the succeeding months t_3, t_4, $\cdots t_8$ in these same proportions, 6, 5, 4, 3, 2, 1; and so on indefinitely, as indicated below:

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}.
X_1·	6	5	4	3	2	1	–	–	–	
X_2.		6	5	4	3	2	1	–	–	
X_3			6	5	4	3	2	1	–	
X_4:				6	5	4	3	2	1.	
X_5:					6	5	4	3	2.	
X_6:						6	5	4	3.	
X_7:							6	5	4.	
X_8:								6	5.	
X_9:									6.	

What, then, is the combined effect, at time t_7, of all the previous X's? The effect of X_1 at time t_7 is $\frac{1}{21} X_1$, as already noted, and as indicated in the above schedule by the figure 1 under t_7. Similarly, the effect of X_2 at this same time t_7 is $\frac{2}{21} X_2$; that of X_3 is $\frac{3}{21} X_3$; and so on.

The total of all these effects combined, at time t_7, is

$$\overline{X}_7 = \frac{X_1 + 2X_2 + 3X_3 + 4X_4 + 5X_5 + 6X_6}{1+2+3+4+5+6}.$$

Note that this total combined effect, at time t_7, of the preceding X_1, X_2, X_3, X_4, X_5, X_6, is called $\overline{X}_7$ and is an average of said preceding X's

Frequently we need not take the trouble to apply the common divisor 21 ($= 1 + 2 + \cdots + 6$), since we are essentially concerned only with the relative magnitudes of the $\overline{X}$'s. We may then prefer to compute merely the numerator of the above fraction which, in contrast to $\overline{X}_7$, may be called ΣX_7; and $\Sigma X_7 = 21\overline{X}_7$, $\Sigma X_8 = 21\overline{X}_8$, etc More generally $\Sigma X_{N+1} = 1X_1 + 2X_2 + 3X_3 + \cdots + NX_N$.

Thus if, in the numerical example above, the values of X_1, X_2, X_3, X_4, X_5, X_6 are respectively 2, 1, 3, 2, 4, 6, then $\Sigma X_7 = 1(2) + 2(1) + 3(3) + 4(2) + 5(4) + 6(6) = 2 + 2 + 9 + 8 + 20 + 36 = 77$ Since $N = 6$, the value of $\overline{X}_7$, the average of the X's corresponding to X_7, is $X_7 = \frac{77}{21} = 3.67$ In the same way ΣX_8, ΣX_9, ΣX_{10}, etc., may be calculated.

The weighted moving total values, ΣX_7, ΣX_8, etc. (or $\overline{X}_7$, $\overline{X}_8$, etc.), are, of course, the values of the estimating index; that is, they are the values of the derived X series which are to be correlated with the Y series, ΣX_7 being paired with the 7th Y value, ΣX_8 being paired with the 8th Y value, and so on. Although still further short cuts are available, we shall not undertake to describe them here.

The reader should realize that the adjustment for the distribution of the lag further reduces the effective number of observations in the derived X series, and the number of degrees of freedom available for measuring σ_r (if such a measure is at all legitimate) is correspondingly reduced. The likelihood of obtaining correlation between the Y series and the derived X series when there is no real relationship is even greater than when the adjustment is made for lag without distributing the lag. Consequently the statistician should be very cautious about drawing conclusions as a result of applying this technique; he should insist upon a strong theoretical argument as well as an abundance of statistical evidence.

Methods of Forecasting

In earlier correlation chapters we have been accustomed to making estimates of the value of one variable from our knowledge of the value of another variable (or variables) and our knowledge concerning the functional relationship between (or among) these variables. These estimates involve the inference that the relationship which has been inferred from the sample is the one that really exists in the population. When correlating time series, however, it is not usually correct to think of the correlated data as constituting a sample from the parent population. Economic relationships are man made, and they change with changes in environment. A forecast may gradually lose its efficacy with the passage of time. Forecasting involves another difficulty also. There is an interval of time between the "cause" X and the "effect" Y. In that interval of time other causes, unknown at the time of the forecast, may intervene, so that the effect may be quite different from that which was originally anticipated. Since this is true, complete reliance cannot be placed on any procedure, statistical or otherwise, for economic forecasting.

Nevertheless, it is the function of a science to make predictions, and, as a practical matter, all rational persons do make forecasts whenever they make any commitment concerning the future. Therefore, any clue that will help us to guess right concerning the future significantly more often than we guess wrong is worth noting. In the following sections we shall summarize some of the methods in vogue at the present time.

Economic rhythm method. In earlier chapters, methods were explained for measuring trend, obtaining periodic patterns, and isolating cycles. In

order to forecast future movements of any series that has been analyzed into these elements, it is necessary to go through the following steps in the order indicated:

(1) Project the trend a number of years into the future. This may be done either freehand or by means of the trend equation. (As mentioned earlier, trend projection is fraught with danger.)

(2) Superimpose on this trend, for a period of perhaps a year or two, an estimate of future cyclical movements, being guided by the past cyclical behavior of this series.

(3) Multiply these estimated monthly cyclical trend values by the appropriate seasonal index numbers.

The heart of this method is step 2. The trend line is supposed to represent the normal growth (or decline) of the series. Then, *assuming* that history repeats itself—that cycles of approximately the same amplitude and duration tend to recur—it is a simple procedure to extend the cyclical design of recent years into the future. This can be done mathematically if desired. Experience shows, however, that cycles in most series are not periodic and that the mathematical extrapolation of cycles is not satisfactory.

In this method, as applied by Roger Babson, areas above and below the normal (or $X-Y$) line enclosed by the cyclical curve are noted.[4] See Chart 255. The $X-Y$ line is constructed by computing the average index number of each complete cycle, placing the figure so obtained at the midpoint in terms of the span of the cycle, and drawing a line through such mid-points for the successive cycles. For uncompleted cycles the line is tentatively extended by inspection. If the area above the line exceeds that below the line, the forecast is depression. As the area below the line approaches in area that above the line, the end of the depression becomes near. However, the depression area cannot be used to forecast the size of the prosperity area. Areas are even carried along from one portion of a cycle to another before being equalized (see areas $G+$ and $G-$); but this makes the practical application of the method very difficult. Another difficulty is that an area can be produced by an increase in either amplitude or duration, whereas the business man is primarily interested in predicting the turning points of a cycle. Finally, there is the objection that the equalization is not accurately accomplished until the completion of each cycle permits drawing in the final trend line.

Although this action-and-reaction concept is borrowed from physics, the

---

[4] See *Babson's Reports*, Special Letter, March 27, 1928; also an undated bulletin "Technical Description of the Babsonchart." Additional information was provided to the authors by the Research Department of *Babson's Reports*.

economic reasoning is that American business has a steady and persistent growth, and that cycles are self-generating. The method is applicable also to the forecasting of occurrences other than business cycles. For instance, there is the belief held by some authorities that minor movements in stock prices tend to last a certain length of time, or until a certain number of shares have been sold.

The Dow system is for forecasting swings in stock prices of shorter duration than the business cycle. The system is based upon a theory of resistance levels. If the market has moved "sidewise" within a narrow price range for several weeks, it means that speculative interests have been either accumulating or distributing stocks. Then, when resistance
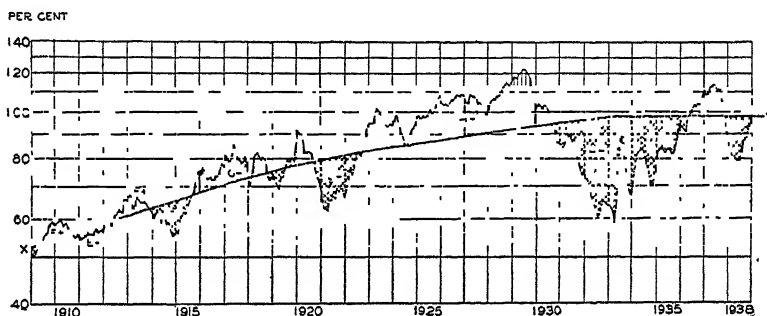
Chart 255. Babson Chart of Physical Volume of Production in the United States (Traced from chart appearing in the January 2, 1939, issue of *Babson's Reports* and published with the Babson Organization's permission. This chart shows a revised trend, or *X-Y* line. In previous charts the trend was extended from the termination of the prolonged depression area which, according to Babson, lasted from the middle of 1930 through most of 1936. This extension, which was shown as a dotted line, sloped upward to the right through 1937 and 1938. The present trend is horizontal from 1933 to date, as is the extension. The effect is to enlarge the 1936–1937 prosperity area somewhat and to make the adjacent depression areas somewhat smaller. Factors supporting the trend revision are briefly described in Babson's December 5, 1938 issue.)

to price change has been broken, if the movement is upward it means the period has been one of "accumulation"; if the "break out" is downward it means that the period has been one of "distribution." If, now, the price movement of the industrials is "confirmed" by one in the same direction by the rails, a sustained movement in that direction is probable (Or, the industrials may confirm the rails.) So long as new highs are made by both groups, a bull market will continue; but if one of the groups persistently fails to confirm the other, the end of the movement may be expected soon. A more complete exposition of this theory is given by Robert Rhea in *The Dow Theory*, Barrons, 1932. While there is some

logic to the Dow theory, it requires exercise of considerable judgment if it is to be used successfully.

Specific historical analogy. Since all cycles are not uniform in amplitude or duration, some forecasters make use of history, not by projecting any fancied economic rhythm into the future but by selecting some specific previous situation which has many of the earmarks of the present, and concluding that what happened in that previous situation will happen in the present one. We quote without comment from Charles G. Dawes' *How Long Prosperity?* (A. N. Marquis Company, Chicago, 1937), p. 24:

> What I present is the unexpected discovery, by a business man in a study without preconceived theory and for the purpose of finding a reasonable basis for the profitable investment of money, of certain important parallels in the last three great depressions in this country of 1873, 1893 and 1929.

Pursuing these parallels, Dawes arrives at the following conclusions (*ibid.*, pp. 38–39):

> 1st. That in the tenth year after the initial stock price collapse in both the 1873 period and the 1893 period, there occurred a stock collapse marking in each case the commencement of a minor business recession.
>
> 2nd. That these minor business recessions (known as those of 1884 and 1904) lasted in the 1884 period approximately two years, and in the 1904 period approximately one year.
>
> 3rd. That prosperous business conditions then ensued.
>
> I predict, therefore, barring wars or inflation of the currency:
>
> 1st. That a high degree of prosperity will maintain in this country into 1939.
>
> 2nd. That beginning in latter part of the year October 1938–October 1939, the tenth year from October 1929, to wit: in the summer or fall of 1939, there will be a stock market collapse.
>
> 3rd. That there will then ensue in the United States a minor recession in business of one or two years.
>
> 4th. That this recession will be followed by a period of prosperity.

The method of specific historical analogy is probably relied upon more heavily by Moody's than by any of the other professional forecasters, although Moody's, like most of the forecasters, does not confine itself to any particular method of forecasting.

Cyclical sequence method. This method, an application of which was dealt with in the section on measurement of lag, is probably the method in greatest favor among forecasters. Sometimes a forecast is based upon the correlation of a dependent variable with one independent variable, as

in the case of industrial production and production of durable consumers goods with a 2-month lag of the former. Again, multiple correlation is used, involving several independent variables.

The forecasting index of Bradford B. Smith is an illustration of this second method. The Smith index attempts to forecast, by one year, changes in the American Telephone and Telegraph Company index. The forecasting index is based on a definite hypothesis: that business is good when money or credit is being used, and bad when money or credit is not forthcoming or for any other reason is not being spent. Consequently, all the series included bear directly on monetary or credit factors. There are four series in the index and, according to Smith, they behave characteristically as follows:

1. *Interest rates.* When interest rates are high, long time borrowing for fixed capital expansion is discouraged; when they are low, such borrowing is encouraged. Such discouragement or encouragement takes about a year to translate itself into changes in business activity. The first series is an average of commercial paper rates and time loan (stock exchange) rates. Seasonal variations prior to 1914 were removed, and the series were expressed in terms of per cent deviation from normal—yields of high-grade long term bonds being regarded as normal. Since a high interest rate forecasts low business activity, the signs of the interest series are reversed before being combined into the index.

2. *Monetary gold plus Federal Reserve bank holdings of United States securities.* Banks tend to be liberal in their loaning policy when they have ample reserves, and they become more strict as their reserves dwindle. Naturally, when gold is imported, bank reserves are built up. Also, a program of United States security buying by Federal Reserve banks builds up bank reserves. These two series, gold and United States securities, are therefore added together and expressed as percentages of trend. The movements of this series are concurrent with those of interest rates, a year ahead of business activity.

3. *Changes in security prices.* When security prices are rising, profits, financed by expanding bank credit, are spent freely. If the speculative fever is high, the high interest rates may retard bond issues and construction work, but they may not appreciably reduce profit taking and profit spending. Likewise, falling security prices bring in their train margin calls by brokers, and the providing of that margin uses up cash which might otherwise have been spent. The security price changes included in the index are those of both stocks and bonds (bonds alone before 1919). The series is "the number of points which the securities would rise during a period of one year if they continued to rise at the same rate which the

trend over the past year would indicate."[5]   This series likewise precedes
business activity by one year.

4. *Amount of new long term bond flotations.*   Both corporate and mu-
nicipal issues are included.   The series is the amount issued during the
12-month period ending with the current month, expressed in terms of
per cent of trend.   The theory involving the use of this series is that new
bond issues mean new capital goods construction.

Weights used in combining these relatives were obtained by a process
of multiple correlation.

Smith gives specific warning that no forecasting index is foolproof.   Rea-
sonable confidence may be had in the forecasts only so long as the economic
relationships and practices persist which suggested the inclusion of the
series used in the index.   Up to the time of the publication of the index,
Smith's forecasts of the American Telephone and Telegraph index agreed
with that index quite as well as the various business indexes that are gen-
erally accepted agreed among themselves.   Smith has not kept his fore-
casting index up to date, however, since he believes it to be applicable
only in the "automatic business economy" which existed at the time he
wrote, but not to be applicable to present conditions.

Usually the data to be correlated when making a forecast are cyclical
relatives; in other instances the data are analyzed in some other fashion.
For instance, Karsten cumulates deviations of car shortages from their
average in order to predict interest rates.[6]   Again, a change in general
business may be predicted by the spread between two series, such as the
spread between prices of raw cotton and cloth, or between imports and
exports.   Often the relationship may be improved if the relative rather
than the absolute difference between two series is taken.   Bradford B.
Smith, it will be remembered, includes the ratio of short term interest
rates to bond yields.

Another illustration of this technique is the relationship between the
hog-corn price ratio and cycles in hog marketing.   According to the United
States Bureau of Agricultural Economics, changes in the relationship of
hog prices to corn prices cause changes in hog production which result in
the hog cycle.   As indicated by Chart 256, a period of greater-than-
average hog-corn price ratios results in an increase in hog marketings a
year or two later, whereas a period of smaller-than-average ratios is fol-
lowed by a decrease in marketings.   The hog cycles as computed by the
Bureau are a 12-month moving average; no adjustment is made for trend.

---

[5] Bradford B. Smith, "A Forecasting Index for Business," *Journal of the American
Statistical Association*, Vol. XXVI, No. 174, June 1931, pp. 115–127.

[6] See Karl G. Karsten, "The Theory of Quadrature in Economics," *Journal of the
American Statistical Association*, Vol. XIX, No. 145, March 1924, p. 14.

Occasionally the forecast may be made not from the turning point of a series, but from its position in respect to some base line. Haney's $\frac{P}{V}$ line (commodity prices ÷ volume of trade) is said to forecast business recovery when it crosses the normal trend on the way up, and depression when it falls below normal.[7] Somewhat akin to this idea is Irving Fisher's technique of using percentage changes in price. In general, the rate of price increase is greatest when the price curve crosses normal on the way up, and the rate of price decrease is greatest when the price curve crosses normal on the way down. Finally, as was illustrated earlier in the chapter,
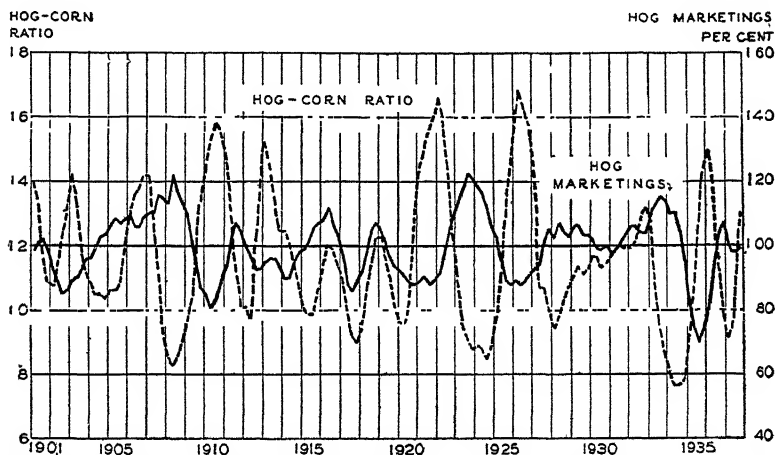
Chart 256. Cycles in Hog-Corn Price Ratio and Hog Marketings, by Quarters, 1901-1937. (Quarterly data adjusted for trend and seasonal variation. Derived from data obtained from the United States Bureau of Agricultural Economics)

the comparison may be between first differences or percentage changes of both series, either with or without adjustment of the original data for trend.

Cautions in regard to the correlation of time series adjusted for lag have already been given. It was pointed out that fortuitous co-variation may appear when there is no economic relationship, and that economic relationships are gradually, and sometimes suddenly, changing. There may be a trend in the lag or it may be that series A will lag at one time and lead at another. Thus, sometimes changes in rates of increase or decrease may be a cause of changes in business activity, and at other times the causal relationship may be reversed. A further qualification of the use-

---

[7] See Lewis H. Haney, *Business Forecasting*, Chapter VIII, Ginn and Co., Boston, 1931. Dr. Haney uses Bradstreet's price index and, for volume of trade, railway freight tonnage adjusted for trend and seasonal variation.

fulness of the method should be added. Series A may precede series B by a certain period of time at revival, but by a different period of time at recession; and it may even be that, while series A may precede series B at recessions, series B may typically precede series A at revivals. A variation of the cyclical sequence method that is simpler than correlation is merely to compute the average number of months by which a given series precedes another at the turning points, revival and recession.[8] These averages may then be applied in forecasting. It is possible also to present a graphic picture of the average relationship between several series in a manner similar to that employed by the Cleveland Trust Company. Chart 257 is based upon two charts published in the Cleveland Trust Company *Business Bulletin* (September 15, 1932, and February 15, 1938)

Through the use of that company's index of business activity for each major depression from 1837 on (13 cycles), a depression index was obtained by averaging the 13 cycles together for the 12th month prior to the lowest point of activity, the 11th month, etc., until an average was obtained for each of the months beginning 12 months prior to the low point and ending 24 months subsequent thereto. The procedure was repeated, using the same time periods for bond prices, stock prices, and commodity prices. The result is shown in section A of Chart 257, which is redrawn from the data of a chart appearing in the *Business Bulletin* of September 15, 1932. Chart 257, section B, is similar to a chart appearing in the *Business Bulletin* of February 15, 1938. The construction of the different series is the same, except that the point of reference is the month during which the business index crossed the normal line on the way down, and the data extend back for the 24 months preceding and the 12 months following this base month. The period of time included in the two charts is of necessity slightly different. According to these charts, a business upturn is preceded by a simultaneous advance in stock and bond prices, and is followed by an upturn in commodity prices. On the downswing the sequence is the same, except that bond prices definitely precede stock prices.

It should be noted that this method of computing lag is not the same as Mitchell's method. The average number of month's lag (Mitchell) is not necessarily the same as the number of months by which the averaged series lags (Cleveland Trust). The correlation method might give still different results.

Cross-cut analysis. This method is based upon the theory that no two cycles are identical, but that like causes always produce like results. All the factors bearing upon a given situation are assembled, and, relying

---

[8] See Wesley C. Mitchell, *Business Cycles, The Problem and Its Setting*, p. 337n., National Bureau of Economic Research, New York, 1927.

upon his knowledge of economic processes, the forecaster concludes whether the situation is favorable or unfavorable.   The Standard Statistics Company relies heavily upon this method.   Although the method is essentially non-statistical,[9] it is possible to develop a statistical technique by
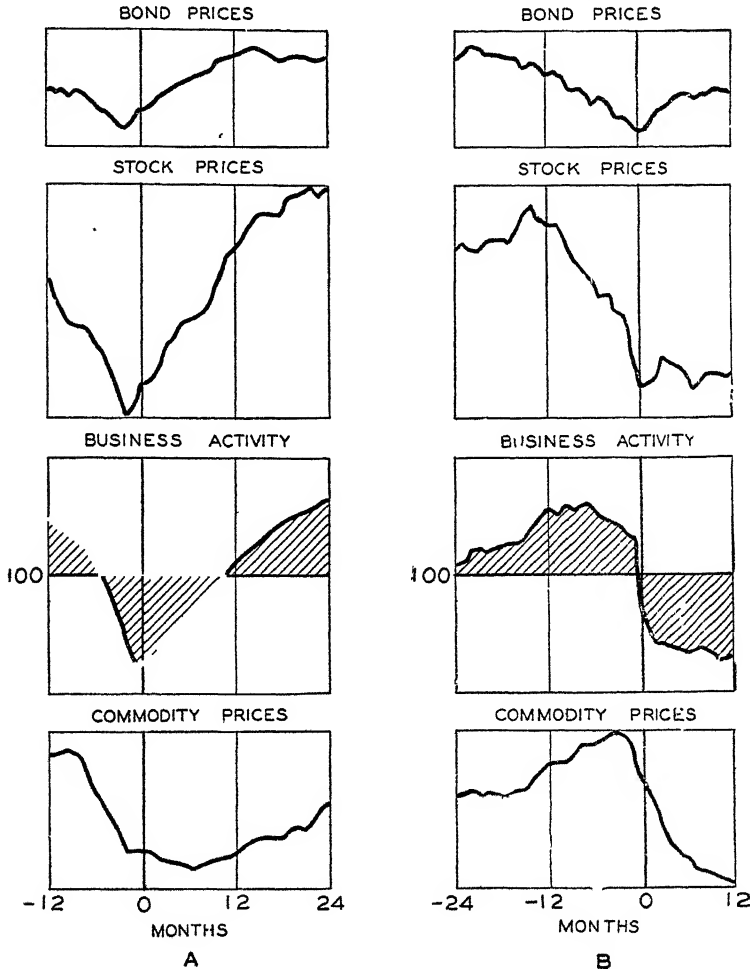
Chart 257.   Some Cyclical Sequences as Determined by the Cleveland Trust Company.   (The vertical scales of the different sections of this chart are not shown, since they are not strictly comparable in an absolute sense.   They were so chosen as to make the amplitude of fluctuation of each part approximately the same.   Based on charts appearing in The Cleveland Trust Company *Business Bulletin*, September 15, 1932 and February 15, 1938.)

[9] Charles O. Hardy and Garfield V. Cox, *Forecasting Business Conditions*, Chapter X. Macmillan Co., New York. 1927.

assigning weights to each factor and then counting the score to see whether the net result is favorable or unfavorable.

The United Business Service has used an interesting type of cross-cut analysis. Instead of marshalling an array of factors bearing upon a given situation, opinions of authorities are assembled and listed each week. A decision is then rendered which is based on the weighted opinions of these authorities modified by the service's own conclusions.

A general caution. Some forecasting devices are built up empirically. A large number of series are compared with the series being forecast, and series which coincide best with that series in a statistical sense are selected, regardless of whether there is any logical cause-and-effect relation. For instance, if it should be found that changes in the anthrax death rate tended regularly to precede turning points in American business, this series might be included in the forecasting index. Since the relationship, if any, was quite accidental, no confidence should be placed in such a forecasting device. It is unwise to forecast upon the basis of statistical data without also obtaining a broad factual knowledge of the changes and developments under way in the field of activity under consideration. A knowledge of underlying economic processes is of basic importance and is essential to the analyst. The statistician who is searching for a magic formula that will enable him to forecast automatically is foredoomed to disappointment. Although it is perhaps a slight exaggeration, it may not be out of place to borrow from Dante and issue the following warning to those who approach the portals of prophecy: "Abandon hope, all ye who enter here."

Selected References

E. C. Bratt: *Business Cycles and Forecasting*, Chapters XIX and XX; Business Publications, Inc., Chicago, 1937.

G. V. Cox: *An Appraisal of American Business Forecasts* (Revised); University of Chicago Press, Chicago, 1930.

F. E. Croxton and D. J. Cowden: *Practical Business Statistics*, Chapter XXI; Prentice-Hall, Inc., New York, 1934.

L H. Haney: *Business Forecasting;* Ginn and Co., Boston, 1931.

C. O. Hardy and G. V. Cox: *Forecasting Business Conditions;* Macmillan Co., New York, 1927.

D. F. Jordan: *Practical Business Forecasting;* Prentice-Hall, Inc., New York, 1927.

W M. Persons· *Forecasting Business Cycles;* John Wiley and Sons, New York, 1931.

W. M. Persons, W. T. Foster, and A. J. Hettinger, Jr.: *Problems of Business Forecasting;* Houghton Mifflin Co., Boston, 1924.

J. R. Rigglemann and I. N. Frisbee: *Business Statistics* (Second Edition), Chapter XVII; McGraw-Hill Book Co., New York, 1938. Describes a number of forecasting methods and services.

J. G. Smith: *Elementary Statistics,* Chapter XV; Henry Holt and Co., New York, 1934.

Carl Snyder. *Business Cycles and Business Measurements,* Chapter XIII; Macmillan Co., New York, 1927.

APPENDICES

APPENDIX A

Selected List of Readily Available Sources

General

(Statistical data covering nearly every field will be found in these general publications.)

Statistical Abstract of the United States. Annual. Bureau of the Census.

The Statesman's Yearbook. Annual. Macmillan and Co., Ltd., London.

The World Almanac and Book of Facts. Annual. New York World-Telegram Co., New York.

Statistical Yearbook of the League of Nations. Annual. League of Nations, Geneva.

Survey of Current Business. Monthly with weekly supplements. Annual supplements also are occasionally issued. Bureau of Foreign and Domestic Commerce.

Federal Reserve Bulletin. Monthly. Board of Governors of the Federal Reserve System.

Standard Trade and Securities: Basic Statistics and monthly bulletins. Standard Statistics Co.

Monthly Bulletin of Statistics of the League of Nations. Monthly. League of Nations, Geneva.

Periodicals, such as:

The Annalist. Weekly. The New York Times Co.

Barrons. Weekly. Barrons Publishing Co.

Business Week. Weekly. McGraw-Hill Publishing Co.

The Magazine of Wall Street. Bi-weekly. The Ticker Publishing Co.

Daily newspapers.

Commodities—Prices, Production, Consumption, Stocks, Exports, and Imports

1. *Census of Agriculture.* Quinquennial. Bureau of the Census.
2. *Yearbook of Agriculture.* Annual, 1894–1935. Department of Agri-

culture. (Since 1935, statistical material has not been included but has been transferred to *Agricultural Statistics.*)

3. *Agricultural Statistics.* Annual. Department of Agriculture.

4. *Crops and Markets.* Monthly. Department of Agriculture.

5. Special studies of the United States Bureau of Agricultural Economics and of the various state agricultural experiment stations

6. *Commerce Yearbook—Vol. I, United States.* Annual Bureau of Foreign and Domestic Commerce. (Discontinued after 1932.)

7. *Foreign Commerce Yearbook.* Annual. Bureau of Foreign and Domestic Commerce. (Formerly *Commerce Yearbook—Vol. II, Foreign Countries.*)

8. *Monthly Summary of Foreign Commerce.* Monthly Bureau of Foreign and Domestic Commerce.

9. *Foreign Commerce and Navigation of the United States.* Annual. Bureau of Foreign and Domestic Commerce.

10. *Foreign Trade of the United States.* Annual. Bureau of Foreign and Domestic Commerce.

11. *The Balance of International Payments of the United States.* Annual. Bureau of Foreign and Domestic Commerce.

12. *Commerce Reports.* Weekly. Bureau of Foreign and Domestic Commerce.

13. *Mineral Resources of the United States.* Annual, 1883–1932. Geological Survey.

14. *Minerals Yearbook.* Annual beginning 1933. (Supersedes *Mineral Resources of the United States.*) Bureau of Mines.

15. *Census of Mines and Quarries.* Decennial. Bureau of the Census.

16 *Census of Manufactures.* Biennial. Bureau of the Census.

17. *Census of Distribution, 1930.* Bureau of the Census.

18. *Census of American Business, 1933.* Bureau of the Census.

19. *Census of Business, 1935.* Bureau of the Census.

20. *Wholesale Prices.* Weekly, monthly, and special bulletins. United States Bureau of Labor Statistics.

21. *Retail Prices.* Bi-weekly, monthly, and special bulletins. United States Bureau of Labor Statistics.

22. *Changes in Cost of Living.* Quarterly. United States Bureau of Labor Statistics.

23. *Record Books of Business Statistics,* issued in conjunction with *Survey of Current Business.* Booklets on textiles, metals and machinery, fuels, automobiles, and rubber issued to date.

24. *Consumer Market Data Handbook.* Annual. Bureau of Foreign and Domestic Commerce.

25. *Sales Management Survey of Buying Power.* Annual Sales Management.

26. *Income in the United States.* Published annually in recent years. Bureau of Foreign and Domestic Commerce.

Financial—Money, Banking, Securities, Interest Rates, Taxation, etc.

1. Bulletins of the individual Federal Reserve banks.
2. Bulletins of various large banks.
3. *Annual Report of the Board of Governors of the Federal Reserve System.*
4. *Annual Report of the Comptroller of the Currency.*
5. Annual reports of the banking departments of various states.
6. *Annual Report of the Federal Deposit Insurance Corporation.* Annual.
7. *Assets and Liabilities of Operating Insured Banks.* Reports on call dates. Federal Deposit Insurance Corporation.
8. *Dun and Bradstreet Monthly Review.* Monthly. Dun & Bradstreet, Inc.
9. *Commercial and Financial Chronicle.* Weekly. William B. Dana Company.
10. *Statistics of Income.* Annual. Bureau of Internal Revenue.
11. *Financial Statistics of States.* Annual. Bureau of the Census.
12. *Financial Statistics of Cities.* Annual. Bureau of the Census.

Business Records of Individual Concerns

1. *Standard Corporation Records.* Daily. Standard Statistics Company.
2. *Moody's Manual of Investments.* (Industrials, railroads, public utilities, governments, banks, etc.) Annual, and bi-weekly bulletins. Moody's Investor's Service.
3. *Poor's Annual.* (Industrials; railroads; public utilities; banks, governments, municipals, investment trusts, real estate, mortgage, finance, and insurance companies) Annual. Poor's Publishing Company.
4. *Insurance Yearbook.* (Life; fire and marine; casualty, surety, and miscellaneous.) Annual. The Spectator Company.
5. Reports of insurance commissioners of various states, especially New York.
6. Annual reports to stockholders of various corporations.

Employment, Wages, and Hours of Labor

1. *Monthly Labor Review.* Monthly. United States Bureau of Labor Statistics.
2. Bulletins of various state bureaus of labor or industrial commissions.
3. Special bulletins of the United States Bureau of Labor Statistics. (Wages and hours of labor, employment and unemployment, productivity of labor, etc.)
4. Special bulletins of the Women's Bureau.
5. *Census of Unemployment* (1930–1931 and 1937). Bureau of the Census.
6. *Census of Occupations.* Decennial. Bureau of the Census.

Miscellaneous

1. *Statistics of Railways in the United States.* Annual. Interstate Commerce Commission.
2. *Report on Value of Water Borne Foreign Commerce.* Annual. United States Maritime Commission. (Until 1936 by United States Shipping Board.)
3. *Annual Report of the Immigration and Naturalization Service.* Department of Justice.
4. *Mortality Statistics.* Annual. Bureau of the Census.
5. *Domestic Commerce.* Monthly. Bureau of Foreign and Domestic Commerce.
6. *Census of Distribution* (1930). Bureau of the Census. (Includes not only retail and wholesale trade, but also distribution of agricultural commodities and census of construction industry, hotels, etc.)
7. *Census of the United States.* Decennial. Bureau of the Census.
8. Various monographs, and annual and special studies of the Bureau of the Census, and the Bureau of Foreign and Domestic Commerce.
9. Bulletins of bureaus of business research of various universities.
10. *Religious Bodies,* 1906, 1916, 1926. Bureau of the Census.

In addition to the above sources, statistical information concerning specific industries may be had from trade papers and trade associations. Lists of trade papers may be found in Ayer and Son's *American Newspaper Annual and Directory* and as an appendix to Thomas' *Register of American Manufacturers* The latter contains also a list of trade associations, found in the appendix entitled "Commercial Organizations." *The Classified List of Trade and Allied Associations and Publications in New York City,* issued by the Chamber of Commerce of the State of New York, lists both trade associations and trade papers.

An appendix to the 1936 Annual Supplement of the *Survey of Current Business* gives sources of data under the following headings:

1. Government departments.
2. Commercial and trade associations.
3. Private organizations.
4. Technical periodicals.

APPENDIX B

Mathematical Appendix

Section IX-1

To prove that $\Sigma x = 0$.

Let $\quad x_1 = X_1 - \overline{X}, x_2 = X_2 - \overline{X}. \quad \cdots, x_N = X_N - \overline{X}.$

Then

$$\Sigma x = \Sigma(X - \overline{X})$$
$$= \Sigma X - N\overline{X}.$$

But $\quad \overline{X} = \dfrac{\Sigma X}{N}$ and $\Sigma X = N\overline{X}.$

Therefore $\quad \Sigma x = N\overline{X} - N\overline{X} = 0.$

Section IX-2

To prove that $\overline{X} = \overline{X}_d + \dfrac{\Sigma d}{N}.$

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_N}{N}.$$

Adding and subtracting $\overline{X}_d$,

$$\overline{X} = \overline{X}_d + \frac{(X_1 - \overline{X}_d) + (X_2 - \overline{X}_d) + \cdots + (X_N - \overline{X}_d)}{N}.$$

But, by definition,

$$d_1 = X_1 - \overline{X}_d, d_2 = X_2 - \overline{X}_d, \cdots, d_N = X_N - \overline{X}_d.$$

Then

$$\overline{X} = \overline{X}_d + \frac{d_1 + d_2 + \cdots + d_N}{N}$$

$$= \overline{X}_d + \frac{\Sigma d}{N}.$$

829

If each item is weighted by its frequency, the expression is

$$\bar{X} = \bar{X}_d + \frac{\Sigma fd}{N}.$$

Section IX-3

To prove that $\bar{X} \geqq G$.

Let X_1 and X_2 be, respectively, the smallest and largest values of a series.

Let $\dfrac{a}{2}$ be the difference between the arithmetic mean and the geometric mean of these values. That is,

$$\frac{X_1 + X_2}{2} = \sqrt{X_1 X_2} + \frac{a}{2},$$

$$X_1 + X_2 = 2\sqrt{X_1 X_2} + a.$$

$$a = X_1 - 2\sqrt{X_1 X_2} + X_2$$

$$= (\sqrt{X_1} - \sqrt{X_2})^2.$$

Therefore $\dfrac{a}{2}$ is either positive or, if $X_1 = X_2$, $\dfrac{a}{2} = 0$ and

$$\frac{X_1 + X_2}{2} \geqq \sqrt{X_1 X_2}.$$

If, now, X_1 and X_2 are each replaced by $\dfrac{X_1 + X_2}{2}$, the value of the arithmetic mean of the entire series is not affected. The value of the geometric mean is, however, *increased* because, as shown above, when $X_1 \neq X_2$, $\dfrac{X_1 + X_2}{2} > \sqrt{X_1 X_2}$ and thus the contribution of $\left(\dfrac{X_1 + X_2}{2}\right)^2$ to the geometric mean exceeds the original contribution of $X_1 X_2$. Continually repeating this process for the smallest and largest remaining values results in a continually increasing value of G which approaches $\bar{X}$.

Section IX-4

To prove that $G \geqq H$.

Let X_1 and X_2 be, respectively, the smallest and largest values of a series.

It was shown in the preceding proof that

$$\frac{X_1 + X_2}{2} \geqq \sqrt{X_1 X_2}.$$

Therefore

$$X_1 + X_2 \geqq 2\sqrt{X_1 X_2}$$

$$\sqrt{X_1 X_2}(X_1 + X_2) \geqq 2(X_1 X_2)$$

$$\sqrt{X_1 X_2} \geqq \frac{2 X_1 X_2}{X_1 + X_2}.$$

But $\dfrac{2 X_1 X_2}{X_1 + X_2} = \dfrac{2}{\dfrac{X_1 + X_2}{X_1 X_2}} = \dfrac{2}{\dfrac{1}{X_1} + \dfrac{1}{X_2}}$, which is the harmonic mean.

If X_1 and X_2 are replaced by their harmonic mean, $\dfrac{2 X_1 X_2}{X_1 + X_2}$, the value of H for the entire series is unchanged. However, the value of G is decreased, for it was shown above that $\sqrt{X_1 X_2} > \dfrac{2 X_1 X_2}{X_1 + X_2}$ when $X_1 \neq X_2$ and thus the contribution of $\left(\dfrac{2 X_1 X_2}{X_1 + X_2}\right)^2$ to the geometric mean would be less than the contribution of $X_1 X_2$. Continually repeating this process for the smallest and largest remaining values results in a continually decreasing value of G which approaches H.

Section X-1

To show that $\quad \sqrt{\dfrac{\Sigma x^2}{N}} = \sqrt{\dfrac{\Sigma d^2}{N} - \left(\dfrac{\Sigma d}{N}\right)^2}.$

Since $\qquad x = X - \overline{X},$

$$\sqrt{\frac{\Sigma x^2}{N}} = \sqrt{\frac{\Sigma(X - \overline{X})^2}{N}}$$

$$= \sqrt{\frac{\Sigma(X^2 - 2X\overline{X} + \overline{X}^2)}{N}}$$

$$= \sqrt{\frac{\Sigma X^2 - 2\overline{X}\Sigma X + N\overline{X}^2}{N}}.$$

But since $\qquad \dfrac{\Sigma X}{N} = \overline{X},$

$$\sqrt{\frac{\Sigma x^2}{N}} = \sqrt{\frac{\Sigma X^2}{N} - 2\overline{X}^2 + \overline{X}^2}$$

$$= \sqrt{\frac{\Sigma X^2}{N} - \overline{X}^2}$$

$$= \sqrt{\frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2}.$$

By definition, $d = X - \bar{X}_d$, or $X = d + \bar{X}_d$.

Therefore·

$$\sqrt{\frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2} = \sqrt{\frac{\Sigma(d + \bar{X}_d)^2}{N} - \left[\frac{\Sigma(d + \bar{X}_d)}{N}\right]^2}$$

$$= \sqrt{\frac{\Sigma(d^2 + 2d\bar{X}_d + \bar{X}_d^2)}{N} - \left(\frac{\Sigma d + N\bar{X}_d}{N}\right)^2}$$

$$= \sqrt{\frac{\Sigma d^2 + 2\bar{X}_d\Sigma d + N\bar{X}_d^2}{N} - \frac{(\Sigma d)^2 + 2N\bar{X}_d\Sigma d + N^2\bar{X}_d^2}{N^2}}$$

$$= \sqrt{\frac{\Sigma d^2}{N} + 2\bar{X}_d\frac{\Sigma d}{N} + \bar{X}_d^2 - \frac{(\Sigma d)^2}{N^2} - 2\bar{X}_d\frac{\Sigma d}{N} - \bar{X}_d^2}$$

$$= \sqrt{\frac{\Sigma d^2}{N} - \left(\frac{\Sigma d}{N}\right)^2}.$$

For a frequency distribution,

$$\sigma = \sqrt{\frac{\Sigma f x^2}{N}}, \text{ and } \sqrt{\frac{\Sigma f x^2}{N}} = \sqrt{\frac{\Sigma f d^2}{N} - \left(\frac{\Sigma f d}{N}\right)^2}.$$

Or, with deviations in terms of class intervals,

$$\sqrt{\frac{\Sigma f x^2}{N}} = i\sqrt{\frac{\Sigma f (d')^2}{N} - \left(\frac{\Sigma f d'}{N}\right)^2}.$$

Section X-2

To show that $\dfrac{Q_1 + Q_3 - 2Q_2}{\dfrac{Q_3 - Q_1}{2}}$ must vary within the limits of ± 2.

If extreme values are added to the upper part of a series, Q_3 moves away from Q_2, and Q_1 moves toward Q_2. As a limiting condition, assume $Q_1 = Q_2$ and, since $Q_1 - Q_2 = 0$, the expression for skewness becomes

$$\frac{Q_3 - Q_2}{(Q_3 - Q_2) \div 2} = +2.$$

Similarly, if extreme values are added to the lower end of a series, Q_1 moves away from Q_2, and Q_3 moves toward Q_2, resulting in a limiting value of -2.

Section X-3

To prove that $\dfrac{\Sigma x^3}{N} = \dfrac{\Sigma d^3}{N} - 3\dfrac{\Sigma d}{N}\dfrac{\Sigma d^2}{N} + 2\left(\dfrac{\Sigma d}{N}\right)^3$

It was shown in Appendix B, section IX–2 that

$$\overline{X} = \overline{X}_d + \frac{\Sigma d}{N}.$$

For any selected X value, say X_1, $\quad x_1 = X_1 - \overline{X} = X_1 - \overline{X}_d - \dfrac{\Sigma d}{N}.$

But $X_1 - \overline{X}_d = d_1$; therefore, $x_1 = d_1 - \dfrac{\Sigma d}{N}.$

Similarly, $\quad x_2 = d_2 - \dfrac{\Sigma d}{N}, \quad x_3 = d_3 - \dfrac{\Sigma d}{N},$ etc.

Thus, $\quad \dfrac{\Sigma x^3}{N} = \dfrac{\Sigma \left(d - \dfrac{\Sigma d}{N}\right)^3}{N},$

$$= \frac{\Sigma\left[d^3 - 3\dfrac{\Sigma d}{N} d^2 + 3\left(\dfrac{\Sigma d}{N}\right)^2 d - \left(\dfrac{\Sigma d}{N}\right)^3 \right]}{N},$$

$$= \frac{\Sigma d^3 - 3\dfrac{\Sigma d}{N}\Sigma d^2 + 3\left(\dfrac{\Sigma d}{N}\right)^2 \Sigma d - N\left(\dfrac{\Sigma d}{N}\right)^3}{N},$$

$$= \frac{\Sigma d^3}{N} - 3\frac{\Sigma d}{N}\frac{\Sigma d^2}{N} + 3\left(\frac{\Sigma d}{N}\right)^2\frac{\Sigma d}{N} - \left(\frac{\Sigma d}{N}\right)^3,$$

$$= \frac{\Sigma d^3}{N} - 3\frac{\Sigma d}{N}\frac{\Sigma d^2}{N} + 3\left(\frac{\Sigma d}{N}\right)^3 - \left(\frac{\Sigma d}{N}\right)^3,$$

$$= \frac{\Sigma d^3}{N} - 3\frac{\Sigma d}{N}\frac{\Sigma d^2}{N} + 2\left(\frac{\Sigma d}{N}\right)^3.$$

Section XI-1

R. A. Fisher sets forth a set of criteria for testing the normality of a distribution. (*Statistical Methods for Research Workers*, Seventh Edition. pp. 54–56 and 74–80.) He computes k_1, k_2, k_3, and k_4, the last three of which are somewhat similar to π_2, π_3, and π_4 except that degrees of freedom are taken into consideration. Thus:

$$k_1 = \frac{\Sigma X}{N}.$$

$$k_2 = \frac{\Sigma x^2}{N - 1}.$$

$$k_3 = \frac{N(\Sigma x^3)}{(N - 1)(N - 2)}.$$

$$k_4 = \frac{N}{(N - 1)\ N - 2)(N - 3)}\left[(N + 1)\Sigma x^4 - 3\frac{N - 1}{N}(\Sigma x^2)^2 \right].$$

(The values of k_2 and k_4 may be corrected for grouping as follows:

$$k_2' = k_2 - \tfrac{1}{12}, \text{ and } k_4' = k_4 + \tfrac{1}{120}$$

No correction is necessary for k_1 and k_3; hence $k_1' = k_1$ and $k_3' = k_3$.)

From the k's, there may be computed the values of

$$g_1 = \frac{k_3}{\sqrt{k_2^3}}, \text{ a measure of skewness,}$$

and

$$g_2 = \frac{k_4}{k_2^2}, \text{ a measure of kurtosis,}$$

which should each be zero for a normal distribution and which are distributed normally for large samples. The variances of g_1 and g_2 are then determined to ascertain whether g_1 and g_2 differ significantly from zero. The expressions are:

$$\sigma_{g_1}^2 = \frac{6N(N-1)}{(N-2)(N+1)(N+3)};$$

$$\sigma_{g_2}^2 = \frac{24N(N-1)^2}{(N-3)(N-2)(N+3)(N+5)}.$$

Section XII-1

To prove that $\sigma_{\bar{x}} = \dfrac{\sigma_P}{\sqrt{N}}$, when P is infinite or very large.

Samples of N items each are drawn at random from a population of P items, as indicated below. There are ${}_P C_N$ such samples.

Item	Sample 1	Sample 2	Sample 3
a	X_{a1}	X_{a2}	X_{a3}
b	X_{b1}	X_{b2}	X_{b3}
c	X_{c1}	X_{c2}	X_{c3}
.	.	.	.
.	.	.	.
.	.	.	.
N	X_{N1}	X_{N2}	X_{N3}

Letting x represent a deviation from the population mean, we have $x_{a1} = X_{a1} - \bar{X}_P, x_{b1} = X_{b1} - \bar{X}_P, \cdots, x_{N1} = X_{N1} - \bar{X}_P, x_{a2} = X_{a2} - \bar{X}_P,$ etc. We shall designate the various items as $\bar{X}_P + x_{a1}, \bar{X}_P + x_{b1}, \cdots, \bar{X}_P + x_{N1}, \bar{X}_P + x_{a2},$ etc.

For sample 1: $\Sigma X_1 = N\bar{X}_P + \Sigma x_1,$

For sample 2: $\Sigma X_2 = N\bar{X}_P + \Sigma x_2,$ etc.

Since adding a constant to (or subtracting a constant from) a series of values does not alter the value of σ, we have

$$\sigma_{\underset{a}{\overset{N}{\Sigma X}}} = \sigma_{\underset{a}{\overset{N}{\Sigma x}}},$$

(where x is a deviation from $\overline{X}_P$ and therefore for any sample $\overset{N}{\underset{a}{\Sigma}} x \neq 0$),

and therefore

$$\sigma^2_{\underset{a}{\overset{N}{\Sigma X}}} = \frac{\overset{_PC_N}{\underset{1}{\Sigma}}\left(\overset{N}{\underset{a}{\Sigma}} x\right)^2}{_PC_N} - \left(\frac{\overset{_PC_N}{\underset{1}{\Sigma}}\overset{N}{\underset{a}{\Sigma}} x}{_PC_N}\right)^2 = \frac{\overset{_PC_N}{\underset{1}{\Sigma}}\left(\overset{N}{\underset{a}{\Sigma}} x\right)^2}{_PC_N},$$

since $\overset{_PC_N}{\underset{1}{\Sigma}}\overset{N}{\underset{a}{\Sigma}} x = \overset{N}{\underset{a}{\Sigma}} x_1 + \overset{N}{\underset{a}{\Sigma}} x_2 + \cdots + \overset{N}{\underset{a}{\Sigma}} x_{_PC_N} = 0$,

and

$$_PC_N \sigma^2_{\underset{a}{\overset{N}{\Sigma X}}} = \overset{_PC_N}{\underset{1}{\Sigma}}\left(\overset{N}{\underset{a}{\Sigma}} x\right)^2 = \overset{_PC_N}{\underset{1}{\Sigma}} (x_a + x_b + x_c + \cdots + x_N)^2.$$

Now for any one sample, for example sample 1,

$$\left(\overset{N}{\underset{a}{\Sigma}} x\right)^2 = (x_a + x_b + x_c + \cdots + x_N)^2$$

$$= x_a^2 + x_a x_b + x_a x_c + \cdots + x_a x_N$$
$$+ x_a x_b + x_b^2 + x_b x_c + \cdots + x_b x_N$$
$$+ x_a x_c + x_b x_c + x_c^2 + \cdots + x_c x_N + \cdots$$
$$+ x_a x_N + x_b x_N + x_c x_N + \cdots + x_N^2$$

$$= \overset{N}{\underset{a}{\Sigma}} x^2 + 2\overset{N}{\underset{a}{\Sigma}} x_i x_j,$$

where $x_i x_j$ represents the product resulting from each combination of two different items. Therefore

$$_PC_N \sigma^2_{\underset{a}{\overset{N}{\Sigma X}}} = \overset{_PC_N}{\underset{1}{\Sigma}}\left(\overset{N}{\underset{a}{\Sigma}} x^2 + 2\overset{N}{\underset{a}{\Sigma}} x_i x_j\right).$$

Since there are $_PC_N$ samples each containing $\frac{N}{P}$ of the population, any

given item (x_i) will occur in $\frac{N}{P}$ of the samples, or $\frac{N}{P} {}_PC_N$ times. Thus each

x^2 will occur $\frac{N}{P} {}_PC_N$ times. Now if a given item (x_i) occurs in $\frac{N}{P}$ of the

samples, a second item (x_j) will occur in $\frac{N-1}{P-1}$ of the samples in which

the first occurs, and both of these items will occur in $\frac{N(N-1)}{P(P-1)}$ of all the

samples, or $\dfrac{N(N-1)}{P(P-1)}{}_{P}C_{N}$ times.   Thus each $x_{i}x_{j}$ will occur $\dfrac{N(N-1)}{P(P-1)}{}_{P}C_{N}$ times.

Therefore

$$
{}_{i}C_{N}\sigma^{2}_{\substack{N\\\Sigma X\\a}} = \frac{N}{P}{}_{P}C_{N}\underset{P}{\Sigma}x^{2} + 2\frac{N(N-1)}{P(P-1)}{}_{P}C_{N}\underset{P}{\Sigma}x_{i}x_{j},
$$

where $\underset{P}{\Sigma}$ indicates a summation over the entire population, and

$$
\sigma^{2}_{\substack{N\\\Sigma X\\a}} = \frac{N}{P}\underset{P}{\Sigma}x^{2} + 2\frac{N(N-1)}{P(P-1)}\underset{P}{\Sigma}x_{i}x_{j}.
$$

Now

$$
(\underset{P}{\Sigma}x)^{2} = \underset{P}{\Sigma}x^{2} + 2\underset{P}{\Sigma}x_{i}x_{j}
$$

$\left[\text{by a development similar to that shown above for } \left(\underset{a}{\overset{N}{\Sigma}}x\right)^{2}\right]$ and

$$
2\underset{P}{\Sigma}x_{i}x_{j} = (\underset{P}{\Sigma}x)^{2} - \underset{P}{\Sigma}x^{2}.
$$

But $\underset{P}{\Sigma}x = 0$; therefore $2\underset{P}{\Sigma}x_{i}x_{j} = -\underset{P}{\Sigma}x^{2}$, and

$$
\sigma^{2}_{\substack{N\\\Sigma X\\a}} = \frac{N}{P}\underset{P}{\Sigma}x^{2} - \frac{N(N-1)}{P(P-1)}\underset{P}{\Sigma}x^{2}
$$

$$
= \frac{N}{P}P\sigma^{2}_{P} - \frac{N(N-1)}{P(P-1)}P\sigma^{2}_{P}
$$

$$
= N\sigma^{2}_{P} - \frac{N(N-1)}{P-1}\sigma^{2}_{P}
$$

$$
= N\sigma^{2}_{P}\left(1 - \frac{N-1}{P-1}\right)
$$

$$
= N\sigma^{2}_{P}\left(\frac{P-1}{P-1} - \frac{N-1}{P-1}\right)
$$

$$
= N\sigma^{2}_{P}\frac{P-N}{P-1}.
$$

$$
\sigma_{\substack{N\\\Sigma X\\a}} = \sqrt{N}\,\sigma_{P}\sqrt{\frac{P-N}{P-1}}.
$$

If each sample contains N items, each deviation of a sample sum from the mean of the sample sums is N times as large as each deviation of a sample mean from the mean of the sample means, each squared deviation of a sample sum is N^{2} times the squared deviation of each sample mean, and the standard deviation of a series of sums is N times the standard deviation of a series of means.   Since all possible samples of size N were

selected, the mean of the sample means $= \bar{X}_P$.   Therefore, dividing each side of the equation by N, we have

$$\sigma_{\bar{x}} = \frac{\sigma_P}{\sqrt{N}}\sqrt{\frac{P-N}{P-1}}.$$

If 1 is negligible in relation to P,

$$\sigma_{\bar{x}} = \frac{\sigma_P}{\sqrt{N}}\sqrt{1 - \frac{N}{P}}.$$

If P is infinite or if P is very large in relation to N, the expression becomes

$$\sigma_{\bar{x}} = \frac{\sigma_P}{\sqrt{N}}.$$

Section XII-2

To prove that $\bar{\sigma} = \sqrt{\dfrac{\Sigma x^2}{N-1}}.$

For any sample composed of items $X_a,\ X_b,\ X_c,\ \cdots,\ X_N,$

$$\sigma^2 = \frac{\overset{N}{\underset{a}{\Sigma}}x^2}{N} = \frac{\overset{N}{\underset{a}{\Sigma}}(X - \bar{X})^2}{N}$$

$$= \frac{\overset{N}{\underset{a}{\Sigma}}X^2 - 2\bar{X}\overset{N}{\underset{a}{\Sigma}}X + N\bar{X}^2}{N}$$

$$= \frac{\overset{N}{\underset{a}{\Sigma}}X^2 - \dfrac{2\left(\overset{N}{\underset{a}{\Sigma}}X\right)^2}{N} + \dfrac{\left(\overset{N}{\underset{a}{\Sigma}}X\right)^2}{N}}{N}$$

$$= \frac{\overset{N}{\underset{a}{\Sigma}}X^2 - \dfrac{\left(\overset{N}{\underset{a}{\Sigma}}X\right)^2}{N}}{N}$$

$$= \frac{N\overset{N}{\underset{a}{\Sigma}}X^2 - \left(\overset{N}{\underset{a}{\Sigma}}X\right)^2}{N^2}.$$

Following the demonstration given in the preceding proof,

$$\left(\overset{N}{\underset{a}{\Sigma}}X\right)^2 = (X_a + X_b + X_c + \cdots + X_N)^2$$

$$= \overset{N}{\underset{a}{\Sigma}}X^2 + 2\overset{N}{\underset{a}{\Sigma}}X_iX_r.$$

Therefore

$$\sigma^2 = \frac{N\sum_a^N X^2 - \left(\sum_a^N X^2 + 2\sum_a^N X_i X_j\right)}{N^2}$$

$$= \frac{N\sum_a^N X^2 - \sum_a^N X^2 - 2\sum_a^N X_i X_j}{N^2}$$

$$= \frac{\sum_a^N X^2(N-1) - 2\sum_a^N X_i X_j}{N^2}.$$

Now there are $_PC_N$ possible samples, each consisting of N items, taken from a population of P items. Therefore, using $\bar{\sigma}^2$ to designate the mean of the variances of all the samples,

$$\overline{[\sigma^2]} = \frac{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_{PC_N}^2}{_PC_N} = \frac{\sum_1^{PC_N}\sigma^2}{_PC_N}$$

$$= \frac{1}{_PC_N}\sum_1^{PC_N}\left[\frac{(N-1)\sum_a^N X^2 - 2\sum_a^N X_i X_j}{N^2}\right].$$

As each sample consists of N items, it contains $\frac{N}{P}$ of the population and any given item (X_i) will occur in $\frac{N}{P}$ of the samples. But, since there are $_PC_N$ samples, each item will occur altogether $\frac{N}{P}\,_PC_N$ times and each X^2 will occur $\frac{N}{P}\,_PC_N$ times. Now, if X_i will occur in $\frac{N}{P}$ of the samples, any other given item (X_j) will occur in $\frac{N-1}{P-1}$ of the samples in which X_i occurs. and X_i and X_j will both occur in $\frac{N(N-1)}{P(P-1)}$ of all the samples. Since there are $_PC_N$ samples, X_i and X_j will both occur in the same sample $\frac{N(N-1)}{P(P-1)}\,_PC_N$ times. Therefore the product X_iX_j will occur $\frac{N(N-1)}{P(P-1)}\,_PC_N$ times. Thus

$$[\overline{\sigma^2}] = \frac{(N-1)\frac{N}{P}\,_PC_N\sum X^2 - 2\frac{N(N-1)}{P(P-1)}\,_PC_N\sum X_iX_j}{_PC_N N^2}$$

$$= \frac{(N-1)\frac{N}{P}\sum X^2 - 2\frac{N(N-1)}{P(P-1)}\sum X_iX_j}{N^2}.$$

It was previously shown that

$$\left(\overset{N}{\underset{a}{\Sigma}}X\right)^2 = \overset{N}{\underset{a}{\Sigma}}X^2 + 2\overset{N}{\underset{a}{\Sigma}}X_i X_j.$$

Similarly,

$$\left(\underset{P}{\Sigma}X\right)^2 = \underset{P}{\Sigma}X^2 + 2\underset{P}{\Sigma}X_i X_j, \text{ and}$$

$$2\underset{P}{\Sigma}X_i X_j = \left(\underset{P}{\Sigma}X\right)^2 - \underset{P}{\Sigma}X^2.$$

Therefore

$$\overline{[\sigma^2]} = \frac{(N-1)\dfrac{N}{P}\underset{P}{\Sigma}X^2 - \dfrac{N(N-1)}{P(P-1)}\left[\left(\underset{P}{\Sigma}X\right)^2 - \underset{P}{\Sigma}X^2\right]}{N^2}$$

$$= \frac{N-1}{N}\frac{\underset{P}{\Sigma}X^2}{P} - \frac{N-1}{N(P-1)}\frac{\left(\underset{P}{\Sigma}X\right)^2}{P} + \frac{N-1}{N(P-1)}\frac{\underset{P}{\Sigma}X^2}{P}$$

$$= \frac{N-1}{N}\frac{\underset{P}{\Sigma}X^2}{P} + \frac{N-1}{N(P-1)}\frac{\underset{P}{\Sigma}X^2}{P} - \frac{P(N-1)}{N(P-1)}\left(\frac{\underset{P}{\Sigma}X}{P}\right)^2$$

$$= \frac{\underset{P}{\Sigma}X^2}{P}\left[\frac{N-1}{N} + \frac{N-1}{N(P-1)}\right] - \left[\frac{P(N-1)}{N(P-1)}\right]\left(\frac{\underset{P}{\Sigma}X}{P}\right)^2.$$

But

$$\frac{N-1}{N} + \frac{N-1}{N(P-1)} = \frac{(N-1)(P-1) + (N-1)}{N(P-1)}$$

$$= \frac{PN - N - P + 1 + N - 1}{N(P-1)}$$

$$= \frac{PN - P}{N(P-1)} = \frac{P(N-1)}{N(P-1)}.$$

Therefore

$$\overline{[\sigma^2]} = \frac{P(N-1)}{N(P-1)}\frac{\underset{P}{\Sigma}X^2}{P} - \frac{P(N-1)}{N(P-1)}\left(\frac{\underset{P}{\Sigma}X}{P}\right)^2$$

$$= \frac{P(N-1)}{N(P-1)}\left[\frac{\underset{P}{\Sigma}X^2}{P} - \left(\frac{\underset{P}{\Sigma}X}{P}\right)^2\right]$$

$$= \frac{P(N-1)}{N(P-1)}\sigma_P^2.$$

As P approaches infinity, $\dfrac{P}{P-1}$ approaches 1 and

$$\overline{[\sigma^2]} = \frac{N-1}{N} \sigma_P^2, \text{ or}$$

$$\sigma_P^2 = \overline{[\sigma^2]} \frac{N}{N-1}.$$

But σ^2 from our sample is the only available estimate of $\overline{[\sigma^2]}$, the mean of the variances of all the samples. Therefore, designating our estimate of σ_P^2 by $\bar{\sigma}^2$,

$$\bar{\sigma}^2 = \sigma^2 \frac{N}{N-1}$$

$$= \frac{\Sigma x^2}{N} \frac{N}{N-1}$$

$$= \frac{\Sigma x^2}{N-1}, \text{ and}$$

$$\bar{\sigma} = \sqrt{\frac{\Sigma x^2}{N-1}}.$$

Section XII-3

To show that $\bar{\sigma} = \sqrt{\dfrac{\Sigma X^2}{N-1} - \dfrac{(\Sigma X)^2}{N(N-1)}}.$

$$\bar{\sigma} = \sqrt{\frac{\Sigma x^2}{N-1}}.$$

$$\bar{\sigma}^2 = \frac{\Sigma x^2}{N-1}$$

$$= \frac{\Sigma(X - \overline{X})^2}{N-1}$$

$$= \frac{\Sigma(X^2 - 2X\overline{X} + \overline{X}^2)}{N-1}$$

$$= \frac{\Sigma X^2 - 2\overline{X}\Sigma X + N\overline{X}^2}{N-1}.$$

But $N\overline{X} = \Sigma X.$

Therefore

$$\bar{\sigma}^2 = \frac{\Sigma X^2 - 2\overline{X}\Sigma X + \overline{X}\Sigma X}{N-1}$$

$$= \frac{\Sigma X^2 - \overline{X}\Sigma X}{N-1}$$

$$= \frac{\Sigma X^2 - \dfrac{(\Sigma X)^2}{N}}{N-1}$$

$$= \frac{\Sigma X^2}{N-1} - \frac{(\Sigma X)^2}{N(N-1)}, \text{ and}$$

$$\tilde{\sigma} = \sqrt{\frac{\Sigma X^2}{N-1} - \frac{(\Sigma X)^2}{N(N-1)}}.$$

Section XII-4

To prove that $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_2}^2 - 2r_{\bar{x}_1\bar{x}_2}\sigma_{\bar{x}_1}\sigma_{\bar{x}_2}}.$

Let $X_{a1}, X_{a2}; X_{b1}, X_{b2}; \cdots ; X_{N1}, X_{N2}$ represent two series of sample values which may be inherently paired or which may have been arbitrarily paired by the process of selection. Thus the standard deviation of the difference between the two series may be computed as follows:

Item	Series 1	Series 2	Difference	Deviation from mean of the differences*
a	X_{a1}	X_{a2}	$X_{a1} - X_{a2}$	$(X_{a1} - X_{a2}) - (\bar{X}_1 - \bar{X}_2)$
b	X_{b1}	X_{b2}	$X_{b1} - X_{b2}$	$(X_{b1} - X_{b2}) - (\bar{X}_1 - \bar{X}_2)$
c	X_{c1}	X_{c2}	$X_{c1} - X_{c2}$	$(X_{c1} - X_{c2}) - (\bar{X}_1 - \bar{X}_2)$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
N	X_{N1}	X_{N2}	$X_{N1} - X_{N2}$	$(X_{N1} - X_{N2}) - (\bar{X}_1 - \bar{X}_2)$

* The mean of the differences is equivalent to the difference between the means, thus

$$\frac{\Sigma(X_1 - X_2)}{N} = \frac{\Sigma X_1 - \Sigma X_2}{N} = \bar{x}_1 - \bar{x}_2$$

$$\sigma_{x_1 - x_2}^2 = \frac{\Sigma[(X_1 - X_2) - (\bar{X}_1 - \bar{X}_2)]^2}{N}$$

$$= \frac{\Sigma[(X_1 - \bar{X}_1) - (X_2 - \bar{X}_2)]^2}{N}$$

$$= \frac{\Sigma(x_1 - x_2)^2}{N}$$

$$= \frac{\Sigma x_1^2}{N} - \frac{2\Sigma x_1 x_2}{N} + \frac{\Sigma x_2^2}{N}$$

But $r_{x_1 x_2} = \dfrac{\Sigma x_1 x_2}{N\sigma_{x_1}\sigma_{x_2}}$ (see Chapter XXII), and therefore

$$\frac{2\Sigma x_1 x_2}{N} = 2r_{x_1 x_2}\sigma_{x_1}\sigma_{x_2}.$$

Also, $\dfrac{\Sigma x_1^2}{N} = \sigma_{x_1}^2$ and $\dfrac{\Sigma x_2^2}{N} = \sigma_{x_2}^2.$

Therefore

$$\sigma^2_{x_1 - x_2} = \sigma^2_{x_1} + \sigma^2_{x_2} - 2r_{x_1 x_2}\sigma_{x_1}\sigma_{x_2} \text{ , and}$$

$$\sigma_{x_1 - x_2} = \sqrt{\sigma^2_{x_1} + \sigma^2_{x_2} - 2r_{x_1 x_2}\sigma_{x_1}\sigma_{x_2}}.$$

When the items of two series are inherently paired and when correlation is present, the above form may be used. However, if the samples of series 1 and series 2 have been drawn independently of each other, $r = 0$, and

$$\sigma_{x_1 - x_2} = \sqrt{\sigma^2_{x_1} + \sigma^2_{x_2}}.$$

If $X_{a1}, X_{a2}; X_{b1}, X_{b2}$; etc. are now taken to be sample means instead of single observations, the expression becomes

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2} - 2r_{\bar{x}_1 \bar{x}_2}\sigma_{\bar{x}_1}\sigma_{\bar{x}_2}}$$

if there is correlation between the paired sample means, and

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2}}$$

if there is no correlation between the paired sample means.

By a procedure similar to the above, it may be shown that

$$\sigma_{x_1 + x_2} = \sqrt{\sigma^2_{x_1} + \sigma^2_{x_2} + 2r_{x_1 x_2}\sigma_{x_1}\sigma_{x_2}}, \text{ and}$$

$$\sigma_{\bar{x}_1 + \bar{x}_2} = \sqrt{\sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2} + 2r_{\bar{x}_1 \bar{x}_2}\sigma_{\bar{x}_1}\sigma_{\bar{x}_2}}.$$

Section XII-5

To show that $\sigma'_{\bar{x}_1 - \bar{x}_2} = \sqrt{\dfrac{(N_1 + N_2)(\Sigma x_1^2 + \Sigma x_2^2)}{N_1 N_2[(N_1 - 1) + (N_2 - 1)]}}.$

The estimated variance of the population from which the first sample was drawn is $\dfrac{\Sigma x_1^2}{N_1 - 1}$, while that from which the second sample was drawn is $\dfrac{\Sigma x_2^2}{N_2 - 1}$. These expressions are each ratios, and it was demonstrated in Chapter VII that ratios based on different N's can be averaged correctly only when properly weighted. The simplest way to accomplish the weighting is to divide the total of the original dividends by the total of the original divisors in both series. Therefore, in averaging the two variances, we total the squared deviations $(\Sigma x_1^2 + \Sigma x_2^2)$ and divide by the total degrees of freedom $[(N_1 - 1) + (N_2 - 1)]$. We have, then,

$$\bar{\sigma}^2_{1+2} = \frac{\Sigma x_1^2 + \Sigma x_2^2}{(N_1 - 1) + (N_2 - 1)}.$$

The expression for the standard error of the difference between two means is

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\bar{\sigma}_1^2}{N_1} + \frac{\bar{\sigma}_2^2}{N_2}}.$$

If, instead of employing two estimates of variance $\bar{\sigma}_1^2$ and $\bar{\sigma}_2^2$, we use a single estimate $\bar{\sigma}_{1+2}^2$, this becomes

$$\sigma'_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\bar{\sigma}_{1+2}^2}{N_1} + \frac{\bar{\sigma}_{1+2}^2}{N_2}}$$

$$= \sqrt{\frac{N_1 \bar{\sigma}_{1+2}^2 + N_2 \bar{\sigma}_{1+2}^2}{N_1 N_2}}$$

$$= \sqrt{\frac{N_1 + N_2}{N_1 N_2} \bar{\sigma}_{1+2}^2}.$$

Substituting for $\bar{\sigma}_{1+2}^2$ gives

$$\sigma'_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{N_1 + N_2}{N_1 N_2} \cdot \frac{\Sigma x_1^2 + \Sigma x_2^2}{(N_1 - 1) + (N_2 - 1)}}$$

$$= \sqrt{\frac{(N_1 + N_2)(\Sigma x_1^2 + \Sigma x_2^2)}{N_1 N_2 [(N_1 - 1) + (N_2 - 1)]}}, \text{ or}$$

$$= \sqrt{\frac{(N_1 + N_2)(\Sigma x_1^2 + \Sigma x_2^2)}{N_1 N_2 (n_1 + n_2)}}.$$

When $N_1 = N_2$, the result is the same whether we employ $\bar{\sigma}_1^2$ and $\bar{\sigma}_2^2$ or $\bar{\sigma}_{1+2}^2$, since

$$\bar{\sigma}_{1+2}^2 = \frac{\Sigma x_1^2 + \Sigma x_2^2}{2N - 2}.$$

$$\sigma'_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{2N}{N^2} \cdot \frac{\Sigma x_1^2 + \Sigma x_2^2}{2N - 2}} = \sqrt{\frac{\Sigma x_1^2 + \Sigma x_2^2}{N(N - 1)}}$$

$$= \sqrt{\frac{\dfrac{\Sigma x_1^2}{N - 1} + \dfrac{\Sigma x_2^2}{N - 1}}{N}} = \sqrt{\frac{\bar{\sigma}_1^2 + \bar{\sigma}_2^2}{N}}$$

$$= \sqrt{\frac{\bar{\sigma}_1^2}{N} + \frac{\bar{\sigma}_2^2}{N}}, \text{ which is } \sigma_{\bar{x}_1 - \bar{x}_2}.$$

Section XII-6

To show that, for a proportionally stratified sample,

$$\sigma^2_{\bar{x} \text{ of a stratified sample}} = \frac{\sigma_P^2}{N} - \frac{\sigma^2_{\text{of strata means}}}{N}.$$

Let there be $a, b, c, \cdots , P_S$ items in a stratum of the population.

Let there be $1, 2, 3, \cdots , m$ strata and let S designate a particular stratum.

A stratified sample of N items is to be selected from a population of P items so that the number of items selected at random from each stratum (N_S) represents the same proportion of the entire sample (N) that the number of items in the stratum (P_S) bears to the population (P). That is, $N_S : N :: P_S : P$.

If we know the true variance of each stratum of the population (σ_S^2), we can, by proper weighting of each stratum variance, obtain the variance in the population around the strata means. Since we are using a stratified sample, and since deviations are measured with reference to the strata means $(\overline{X}_S)$ rather than with reference to the population mean $(\overline{X}_P)$, we shall designate this measure as $\sigma_{P'}^2$ in order to distinguish it from σ_P^2 (the population variance in which each deviation is measured from the population mean). We have then

$$\sigma_{P'}^2 = \frac{\overset{m}{\underset{1}{\Sigma}} \overset{P_S}{\underset{a}{\Sigma}} (X - \overline{X}_S)^2}{\overset{m}{\underset{1}{\Sigma}} P_S} = \frac{\overset{m}{\underset{1}{\Sigma}} P_S \sigma_S^2}{P}$$

Since the general expression for the variance of the mean is $\sigma_{\overline{x}}^2 = \dfrac{\sigma_P^2}{N}$, we may write

$$\sigma_{\overline{x}}^2 \text{ of a stratified sample} = \frac{\sigma_{P'}^2}{N} = \frac{\overset{m}{\underset{1}{\Sigma}} P_S \sigma_S^2}{P} \div N.$$

Now $\sigma_S^2 = \dfrac{\overset{P_S}{\underset{a}{\Sigma}} \left(X - \overline{X}_S \right)^2}{P_S}$, and

$$\overset{P_S}{\underset{a}{\Sigma}} (X - \overline{X}_S)^2 = \overset{P_S}{\underset{a}{\Sigma}} (X - \overline{X}_S - \overline{X}_P + \overline{X}_P)^2$$

$$= \overset{P_S}{\underset{a}{\Sigma}} \left[(X - \overline{X}_P) - (\overline{X}_S - \overline{X}_P) \right]^2$$

$$= \overset{P_S}{\underset{a}{\Sigma}} \left[(X - \overline{X}_P)^2 - 2(X - \overline{X}_P)(\overline{X}_S - \overline{X}_P) + (\overline{X}_S - \overline{X}_P)^2 \right]$$

$$= \overset{P_S}{\underset{a}{\Sigma}} (X - \overline{X}_P)^2 - 2(\overline{X}_S - \overline{X}_P) \overset{P_S}{\underset{a}{\Sigma}} (X - \overline{X}_P) + P_S(\overline{X}_S - \overline{X}_P)^2$$

$$= \overset{P_S}{\underset{a}{\Sigma}} (X - \overline{X}_P)^2 - 2(\overline{X}_S - \overline{X}_P)(\overset{P_S}{\underset{a}{\Sigma}} X - P_S \overline{X}_P) + P_S(\overline{X}_S - \overline{X}_P)^2$$

$$= \overset{P_s}{\underset{a}{\Sigma}}(X - \overline{X}_P)^2 - 2(\overline{X}_S - \overline{X}_P)(P_S\overline{X}_S - P_S\overline{X}_P) \\ + P_S(\overline{X}_S - \overline{X}_P)^2$$

$$= \overset{P_s}{\underset{a}{\Sigma}}(X - \overline{X}_P)^2 - 2P_S(\overline{X}_S - \overline{X}_P)^2 + P_S(\overline{X}_S - \overline{X}_P)^2$$

$$= \overset{P_s}{\underset{a}{\Sigma}}(X - \overline{X}_P)^2 - P_S(\overline{X}_S - \overline{X}_P)^2$$

Therefore

$$\sigma_S^2 = \frac{\overset{P_s}{\underset{a}{\Sigma}}(X - \overline{X}_P)^2}{P_S} - (\overline{X}_S - \overline{X}_P)^2.$$

Now

$$\sigma_{P'}^2 = \frac{\overset{m}{\underset{1}{\Sigma}}P_S\sigma_S^2}{P} = \frac{\overset{m}{\underset{1}{\Sigma}}P_S\left[\dfrac{\overset{P_s}{\underset{a}{\Sigma}}(X - \overline{X}_P)^2}{P_S} - (\overline{X}_S - \overline{X}_P)^2\right]}{P}, \text{ and}$$

$$\sigma_{\overline{X} \text{ of a stratified sample}}^2 = \left\{\frac{\overset{m}{\underset{1}{\Sigma}}P_S\left[\dfrac{\overset{P_s}{\underset{a}{\Sigma}}(X - \overline{X}_P)^2}{P_S} - (\overline{X}_S - \overline{X}_P)^2\right]}{P}\right\} \div N$$

$$= \left\{\frac{\overset{m}{\underset{1}{\Sigma}}P_S\left[\dfrac{\overset{P_s}{\underset{a}{\Sigma}}(X - \overline{X}_P)^2}{P_S}\right]}{P} - \frac{\overset{m}{\underset{1}{\Sigma}}P_S(\overline{X}_S - \overline{X}_P)^2}{P}\right\} \div N$$

$$= \left[\frac{\overset{m\,P_s}{\underset{1\ a}{\Sigma\Sigma}}(X - \overline{X}_P)^2}{P} - \frac{\overset{m}{\underset{1}{\Sigma}}P_S(\overline{X}_S - \overline{X}_P)^2}{P}\right] \div N$$

$$= \frac{\sigma_P^2}{N} - \frac{\sigma_{\text{of strata means}}^2}{N},$$

where σ_P^2 is the population variance computed in the usual fashion with reference to the population mean and $\sigma_{\text{of strata means}}^2$ is the weighted variance of the strata means.

If the population (P) is finite and the sample (N) is large in relation to the population, we must refer to the expression given in section XII–1:

$$\sigma_{\overline{x}} = \frac{\sigma_P}{\sqrt{N}}\sqrt{\frac{P - N}{P - 1}}.$$

Squaring gives

$$\sigma_{\bar{x}}^2 = \frac{\sigma_{\bar{p}}^2}{N} \cdot \frac{P - N}{P - 1}.$$

For a stratified sample this becomes

$$\sigma_{\bar{x} \text{ of a stratified sample}}^2 = \left(\frac{\sum\limits_1^m P_S \sigma_S^2}{P} \div N \right) \left(\frac{P - N}{P - 1} \right) = \left(\frac{\sigma_P^2}{N} - \frac{\sigma_{\text{of strata means}}^2}{N} \right) \left(\frac{P - N}{P - 1} \right).$$

Section XIII-1

To prove that

$$\chi^2 = \frac{\left(a - \frac{p}{q} b \right)^2}{\frac{p}{q} N}.$$

For the data of 40 first cousins, the value of χ^2 was found to be .40 by the usual procedure:

Sex	Observed f	Expected ratio 1 : 1 fc	$f - fc$	$(f - fc)^2$	$\dfrac{(f - fc)^2}{fc}$
Male	22	20	+2	4	.20
Female	18	20	−2	4	.20
Total. ..	40	40	.	.	.40

Using a, b, p, q, and N, as in Chapter XIII, we have from the above

$$\chi^2 = \frac{(a - pN)^2}{pN} + \frac{(b - qN)^2}{qN}$$

$$= \frac{a^2 - 2apN + p^2N^2}{pN} + \frac{b^2 - 2bqN + q^2N^2}{qN}$$

$$= \frac{qa^2 - 2apqN + p^2qN^2 + pb^2 - 2bpqN + q^2pN^2}{Npq}$$

$$= \frac{qa^2 + pb^2 + p^2qN^2 + q^2pN^2 - 2pqN(a + b)}{Npq}.$$

Now $a + b = N$; hence

$$\chi^2 = \frac{qa^2 + pb^2 + p^2qN^2 + q^2pN^2 - 2pqN^2}{Npq}$$

$$= \frac{qa^2 + pb^2 + pqN^2(p + q - 2)}{Npq}.$$

But $p + q = 1$, and $p + q - 2 = -1$; therefore

$$\chi^2 = \frac{qa^2 + pb^2 - pqN^2}{Npq}$$

$$= \frac{qa^2 + pb^2 - pq(a + b)^2}{Npq}$$

$$= \frac{qa^2 + pb^2 - (pqa^2 + 2pqab + pqb^2)}{Npq}.$$

Now $p = 1 - q$, and $q = 1 - p$; therefore

$$\chi^2 = \frac{qa^2 + pb^2 - [(1 - q)qa^2 + 2pqab + (1 - p)pb^2]}{Npq}$$

$$= \frac{qa^2 + pb^2 - (qa^2 - q^2a^2 + 2pqab + pb^2 - p^2b^2)}{Npq}$$

$$= \frac{q^2a^2 - 2pqab + p^2b^2}{Npq}$$

$$= \frac{(qa - pb)^2}{Npq}.$$

Dividing by $\dfrac{q^2}{q^2}$ gives

$$\chi^2 = \frac{\left(a - \dfrac{p}{q}b\right)^2}{\dfrac{p}{q}N}.$$

Section XIII-2

Derivation of Formulae Used for Computing Total Variation, Variation Within Columns, and Variation Between Columns

In the following developments

N_K represents the number of items in a column,
m the number of columns, and
N the number of items in all columns.

$\displaystyle\sum_1^{N_K}$ refers to a summation of items 1 to N_K in a column,

$\displaystyle\sum_1^{m}$ indicates a summation of values for columns 1 to m, and

$\displaystyle\sum_1^{N} = \Sigma$ designates a summation of all items.

A. The deviation of an item in a column from the grand mean $(Y - \overline{Y})$ may be broken into two parts: first, the deviation of the item from the column mean $(Y - \overline{Y}_1)$; and second, the deviation of the column mean from the grand mean $(\overline{Y}_1 - \overline{Y})$. Thus, for an item in the first column,

$$(Y - \overline{Y}) = (Y - \overline{Y}_1) + (\overline{Y}_1 - \overline{Y}).$$

As a measure of total variation, the deviations $(Y - \overline{Y})$ are to be squared and summed. Squaring and summing, first for a single column, we have

$$\sum_{1}^{N_K} (Y - \overline{Y})^2 = \sum_{1}^{N_K} \left[(Y - \overline{Y}_1) + (\overline{Y}_1 - \overline{Y}) \right]^2$$

$$= \sum_{1}^{N_K} \left[(Y - \overline{Y}_1)^2 + 2(Y - \overline{Y}_1)(\overline{Y}_1 - \overline{Y}) + (\overline{Y}_1 - \overline{Y})^2 \right]$$

$$= \sum_{1}^{N_K} (Y - \overline{Y}_1)^2 + 2(\overline{Y}_1 - \overline{Y}) \sum_{1}^{N_K} (Y - \overline{Y}_1) + N_1(\overline{Y}_1 - \overline{Y})^2.$$

Now the summation of the deviations of the Y values of the column from the column mean $\overline{Y}_1$ equals zero; that is, $\sum_{1}^{N_K} (Y - \overline{Y}_1) = 0$, and therefore

$$\sum_{1}^{N_K} (Y - \overline{Y})^2 = \sum_{1}^{N_K} (Y - \overline{Y}_1)^2 + N_1(\overline{Y}_1 - \overline{Y})^2$$

and similarly for all other columns.

Summing the preceding expression for all columns gives

$$\sum_{1}^{m} \left[\sum_{1}^{N_K} (Y - \overline{Y})^2 \right] = \sum_{1}^{m} \left[\sum_{1}^{N_K} (Y - \overline{Y}_K)^2 \right] + \sum_{1}^{m} \left[N_K(\overline{Y}_K - \overline{Y})^2 \right], \text{ or}$$

$$\sum_{1}^{N} (Y - \overline{Y})^2 = \sum_{1}^{m} \left[\sum_{1}^{N_K} (Y - \overline{Y}_K)^2 \right] + \sum_{1}^{m} \left[N_K(\overline{Y}_K - \overline{Y})^2 \right].$$

It is apparent that

$$\sum_{1}^{N} (Y - \overline{Y})^2 = \text{total variation,}$$

$$\sum_{1}^{m} \left[\sum_{1}^{N_K} (Y - \overline{Y}_K)^2 \right] = \text{variation within columns, and}$$

$$\sum_{1}^{m} \left[N_K(\overline{Y}_K - \overline{Y})^2 \right] = \text{variation between columns.}$$

B. For purposes of computation each of the three above expressions may be simplified:

(1) *Total variation,* $\overset{N}{\underset{1}{\Sigma}}(Y - \overline{Y})^2$ or $\Sigma(Y - \overline{Y})^2$. This is the numerator of the expression previously used in computing σ^2.

$$\begin{aligned}
\Sigma(Y - \overline{Y})^2 &= \Sigma(Y^2 - 2Y\overline{Y} + \overline{Y}^2) \\
&= \Sigma Y^2 - 2\overline{Y}\Sigma Y + N\overline{Y}^2 \\
&= \Sigma Y^2 - 2\overline{Y}\Sigma Y + \overline{Y}\Sigma Y \\
&= \Sigma Y^2 - \overline{Y}\Sigma Y \text{ (This form is used in chap-} \\
&\qquad\qquad\qquad\qquad \text{ters on correlation.)} \\
&= \Sigma Y^2 - \frac{(\Sigma Y)^2}{N}.
\end{aligned}$$

(2) *Variation within columns,* $\overset{m}{\underset{1}{\Sigma}}\left[\overset{N_K}{\underset{1}{\Sigma}}(Y - \overline{Y}_K)^2\right]$. This expression says: "For each column, sum the squared deviations from the mean of that column; then sum these totals for all columns." For the first column,

$$\begin{aligned}
\overset{N_K}{\underset{1}{\Sigma}}(Y - \overline{Y}_1)^2 &= \overset{N_K}{\underset{1}{\Sigma}}(Y^2 - 2Y\overline{Y}_1 + \overline{Y}_1^2) \\
&= \overset{N_K}{\underset{1}{\Sigma}}Y^2 - 2\overline{Y}_1\overset{N_K}{\underset{1}{\Sigma}}Y + N_1\overline{Y}_1^2 \\
&= \overset{N_K}{\underset{1}{\Sigma}}Y^2 - 2\overline{Y}_1\overset{N_K}{\underset{1}{\Sigma}}Y + \overline{Y}_1\overset{N_K}{\underset{1}{\Sigma}}Y \\
&= \overset{N_K}{\underset{1}{\Sigma}}Y^2 - \overline{Y}_1\overset{N_K}{\underset{1}{\Sigma}}Y.
\end{aligned}$$

Summing this last expression for all columns gives

$$\overset{m}{\underset{1}{\Sigma}}\left[\overset{N_K}{\underset{1}{\Sigma}}(Y - \overline{Y}_K)^2\right] = \Sigma Y^2 - \overset{m}{\underset{1}{\Sigma}}\left(\overline{Y}_K\overset{N_K}{\underset{1}{\Sigma}}Y\right) \text{ (This form is used in Chapter XXIII.)}$$

$$= \Sigma Y^2 - \overset{m}{\underset{1}{\Sigma}}\left[\frac{\left(\overset{N_K}{\underset{1}{\Sigma}}Y\right)^2}{N_K}\right].$$

If $N_1 = N_2 = \cdots = N_m$, the expression becomes

$$\overset{m}{\underset{1}{\Sigma}}\left[\overset{N_K}{\underset{1}{\Sigma}}(Y - \overline{Y}_K)^2\right] = \Sigma Y^2 - \frac{\overset{m}{\underset{1}{\Sigma}}\left(\overset{N_K}{\underset{1}{\Sigma}}Y\right)^2}{N_K}.$$

(3) *Variation between columns,* $\overset{m}{\underset{1}{\Sigma}}\left[N_K(\overline{Y}_K - \overline{Y})^2\right]$. This expression

says: "For each column, square the deviation of the column mean from the grand mean, multiply by the number of items in the column, and sum these products for all columns."

$$\sum_{1}^{m}\left[N_K(\bar{Y}_K - \bar{Y})^2\right] = \sum_{1}^{m}\left[N_K(\bar{Y}_K^2 - 2\bar{Y}_K\bar{Y} + \bar{Y}^2)\right]$$

$$= \sum_{1}^{m}(N_K\bar{Y}_K^2 - 2N_K\bar{Y}_K\bar{Y} + N_K\bar{Y}^2)$$

$$= \sum_{1}^{m}(N_K\bar{Y}_K^2) - 2\bar{Y}\sum_{1}^{m}(N_K\bar{Y}_K) + \sum_{1}^{m}(N_K\bar{Y}^2).$$

But
$$\sum_{1}^{m}(N_K\bar{Y}_K^2) = \sum_{1}^{m}\left(\bar{Y}_K\sum_{1}^{N_K}Y\right),$$

$$\sum_{1}^{m}(N_K\bar{Y}_K) = \sum_{1}^{m}\left(\sum_{1}^{N_K}Y\right) = \sum Y, \text{ and}$$

$$\sum_{1}^{m}(N_K\bar{Y}^2) = N\bar{Y}^2 = \bar{Y}\sum Y.$$

Therefore
$$\sum_{1}^{m}\left[N_K(\bar{Y}_K - \bar{Y})^2\right] = \sum_{1}^{m}\left(\bar{Y}_K\sum_{1}^{N_K}Y\right) - 2\bar{Y}\sum Y + \bar{Y}\sum Y$$

$$= \sum_{1}^{m}\left(\bar{Y}_K\sum_{1}^{N_K}Y\right) - \bar{Y}\sum Y \quad \text{(This form is used in Chapter XXIII.)}$$

$$= \sum_{1}^{m}\left[\frac{\left(\sum_{1}^{N_K}Y\right)^2}{N_K}\right] - \frac{(\sum Y)^2}{N}.$$

If $N_1 = N_2 = \cdots = N_m$, the expression becomes

$$\sum_{1}^{m}\left[N_K(\bar{Y}_K - \bar{Y})^2\right] = \frac{\sum_{1}^{m}\left(\sum_{1}^{N_K}Y\right)^2}{N_K} - \frac{(\sum Y)^2}{N}.$$

Section XV-1

Derivation of Normal Equations for Straight Line

If Y_C is a trend or computed value, $Y - Y_C$ is a deviation from trend. To satisfy the least-squares criterion, $\sum(Y - Y_C)^2$ must be at a minimum. Since the straight line equation type is $Y_C = a + bX$,

$$\sum(Y - Y_c)^2 = \sum[Y - (a + bX)]^2 = \sum(Y - a - bX)^2.$$

Expanding, this expression becomes

$$\Sigma Y^2 - 2a\Sigma Y - 2b\Sigma XY + Na^2 + 2ab\Sigma X + b^2\Sigma X^2.$$

Calling this expression ϕ, and taking the partial derivative with respect to a, we have

$$\frac{\delta\phi}{\delta a} = -2\Sigma Y + 2Na + 2b\Sigma X.$$

The value of a curve is at a minimum (or maximum) when its slope is zero. Therefore, setting the partial derivative equal to zero, we have

$$-2\Sigma Y + 2Na + 2b\Sigma X = 0,$$
$$\Sigma Y = Na + b\Sigma X, \text{ which is normal equation I.}$$

Differentiating with respect to b:

$$\frac{\delta\phi}{\delta b} = -2\Sigma XY + 2a\Sigma X + 2b\Sigma X^2.$$

Setting the partial derivative equal to zero:

$$-2XY + 2aX + 2bX^2 = 0, \text{ or}$$
$$\Sigma XY = a\Sigma X + b\Sigma X^2, \text{ which is normal equation II.}$$

Section XV-2

The Least Squares Criterion

The following discussion assumes that the distribution of chance errors follows the normal curve, and that the best central value from which to measure such accidental deviations is therefore that value which makes it most probable that the deviations are distributed normally.

Let a series of such deviations, or errors, and the interval within which they fall be designated by the following symbols:

x_1 is an item falling at the mid-point of a very small interval, Δx_1;
x_2 " " " " " " " " " " " " Δx_2;
.   .
.   .
.   .
x_N " " " " " " " " " " " " Δx_N.

Now the probability that a deviation will fall within a certain interval is

$$P = \frac{\text{Area of frequency curve within boundaries of that interval}}{\text{Area of entire frequency curve}}.$$

Thus the probability of obtaining an error x_1 which falls within the interval Δx_1 is approximately the ratio of the area of a rectangle, with base of

Δx_1 and height the ordinate at the mid-point of the interval, to the area of the entire frequency curve.

If this curve is the normal curve, this probability is

$$\frac{i}{\sigma\sqrt{2\pi}} e^{-\frac{x_1^2}{2\sigma^2}} \Delta x_1,$$

since the expression for the ordinate of a normal curve as a ratio to the

entire number of frequencies is $Y_C = \dfrac{i}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}.$

The probability of obtaining errors x_2, x_3, etc., falling within specified intervals is similarly obtained.

The probability that several independent events will occur is the product of the individual probabilities of the separate events. Therefore the probability that the particular *set* of errors will occur which we have assumed (that is, a normal distribution of errors) is as follows:

$$P = \left(\frac{i}{\sigma\sqrt{2\pi}} e^{-\frac{x_1^2}{2\sigma^2}} \Delta x_1\right) \times \left(\frac{i}{\sigma\sqrt{2\pi}} e^{-\frac{x_2^2}{2\sigma^2}} \Delta x_2\right)$$

$$\times \cdots \times \left(\frac{i}{\sigma\sqrt{2\pi}} e^{-\frac{x_N^2}{2\sigma^2}} \Delta x_N\right)$$

$$= \frac{i^N}{\sigma^N 2\pi^{\frac{N}{2}}} e^{-\frac{x_1^2 + x_2^2 + \cdots + x_N^2}{2\sigma^2}} \times \Delta x_1 \times \Delta x_2 \times \cdots \times \Delta x_N.$$

Since any number raised to a negative power will be greatest when that exponent is least, P is greatest when $x_1^2 + x_2^2 + \cdots + x_N^2$ is least. Therefore the probability that accidental deviations from some central value will follow the normal curve is greatest when the sum of the squared deviations from that central value is at a minimum.

Section XVI-1

Derivation of Equations for Fitting Growth Curve $Y_C = k + ab^X$

Designating by n the number of years in each third of the data, the first equation (see equation I, p. 443) is:

$$\Sigma_1 Y = nk + a + ab + ab^2 + ab^3 + \cdots + ab^{(n-1)}$$

$$= nk + a[1 + b + b^2 + b^3 + \cdots + b^{(n-1)}].$$

If now the expression inside the brackets be multiplied by $\dfrac{b-1}{b-1}$, we have

$$\frac{[1 + b + b^2 + b^3 + \cdots + b^{(n-1)}](b - 1)}{b - 1} \quad \cdots \cdots \cdots \cdots \cdots \cdots \cdots (1)$$

$$= \frac{b + b^2 + b^3 + \cdots + b^{(n-1)} + b^n - 1 - b - b^2 - b^3 - \cdots - b^{(n-1)}}{b - 1} \quad \cdots (2)$$

$$= \frac{b^n - 1}{b - 1}.$$

The fourth term shown in the numerator of expression (2) is $b^{(n-1)}$. This follows from the fact that the next to the last term within the brackets of expression (1) may also be designated as $b^{(n-2)}$; and $b^{(n-2)} \times b = b^{(n-1)}$. All three equations are obtained in a similar fashion. They are:

$$\text{I.} \quad \Sigma_1 Y = nk + a\left(\frac{b^n - 1}{b - 1}\right).$$

$$\text{II.} \quad \Sigma_2 Y = nk + ab^n\left(\frac{b^n - 1}{b - 1}\right).$$

$$\text{III.} \quad \Sigma_3 Y = nk + ab^{2n}\left(\frac{b^n - 1}{b - 1}\right).$$

Equations A, B, and C now are:

$$\text{A.} \quad \Sigma_2 Y - \Sigma_1 Y = a\left(\frac{b^n - 1}{b - 1}\right)(b^n - 1) = a\frac{(b^n - 1)^2}{b - 1}.$$

$$\text{B.} \quad \Sigma_3 Y - \Sigma_2 Y = ab^n\frac{(b^n - 1)^2}{b - 1}.$$

$$\text{C.} \quad \frac{\Sigma_3 Y - \Sigma_2 Y}{\Sigma_2 Y - \Sigma_1 Y} = ab^n\frac{(b^n - 1)^2}{b - 1} \div a\frac{(b^n - 1)^2}{b - 1} = b^n.$$

Therefore $b = \sqrt[n]{\dfrac{\Sigma_3 Y - \Sigma_2 Y}{\Sigma_2 Y - \Sigma_1 Y}}.$

Equation A gives us the formula for a:

$$\Sigma_2 Y - \Sigma_1 Y = a\left(\frac{b^n - 1}{b - 1}\right)^2.$$

$$a = \Sigma_2 Y - \Sigma_1 Y\frac{b - 1}{(b^n - 1)^2}.$$

From equation I we find:

$$\Sigma_1 Y = nk + a\frac{b^n - 1}{b - 1}.$$

$$k = \frac{1}{n}\left[\Sigma_1 Y - \left(\frac{b^n - 1}{b - 1}\right)a\right].$$

Section XXII-1

To prove that $\overline{Y}_C = \overline{Y}$.

$$Y_C = a + bX.$$
$$\Sigma Y_C = \Sigma(a + bX)$$
$$= Na + b\Sigma X.$$

But $Na + b\Sigma X = \Sigma Y$ (Normal equation I).

Therefore $\Sigma Y_C = \Sigma Y.$ .. .. (1)

$$\frac{\Sigma Y_C}{N} = \frac{\Sigma Y}{N}, \text{ and}$$

$$\overline{Y}_C = \overline{Y} \quad ....(2)$$

To prove that $\Sigma Y_C^2 = a\Sigma Y + b\Sigma XY$.

$$\Sigma Y_C^2 = \Sigma(a + bX)^2$$
$$= \Sigma(a^2 + 2abX + b^2X^2)$$
$$= Na^2 + 2ab\Sigma X + b^2\Sigma X^2$$
$$= a(Na + b\Sigma X) + b(a\Sigma X + b\Sigma X^2).$$

But $Na + b\Sigma X = \Sigma Y$ (Normal equation I), and

$a\Sigma X + b\Sigma X^2 = \Sigma XY$ (Normal equation II).

Therefore

$$\Sigma Y_C^2 = a\Sigma Y + b\Sigma XY \quad ... . \quad (3)$$

To prove that $\Sigma y_C^2 = \Sigma Y_C^2 - \overline{Y}\Sigma Y$ or $(a\Sigma Y + b\Sigma XY) - \overline{Y}\Sigma Y$.

It has been shown (in Appendix B, section XIII-2) that

$$\Sigma y^2 = \Sigma Y^2 - \overline{Y}\Sigma Y.$$

Similarly it is true that $\Sigma y_C^2 = \Sigma Y_C^2 - \overline{Y}_C\Sigma Y_C$.

But $\overline{Y}_C = \overline{Y}$ (equation 2) and $\Sigma Y_C = \Sigma Y$ (equation 1).

Therefore $\Sigma y_C^2 = \Sigma Y_C^2 - \overline{Y}\Sigma Y$ . ...(4)

However, since $\Sigma Y_C^2 = a\Sigma Y + b\Sigma XY$ (equation 3), it follows that

$$\Sigma y_C^2 = (a\Sigma Y + b\Sigma XY) - \overline{Y}\Sigma Y \quad .... \quad ..(5)$$

To prove that $\Sigma y_S^2 = \Sigma Y^2 - \Sigma Y_C^2$ or $\Sigma Y^2 - (a\Sigma Y + b\Sigma XY)$.

$$\Sigma y_S^2 = \Sigma(Y - Y_C)^2$$
$$= \Sigma Y^2 - 2\Sigma YY_C + \Sigma Y_C^2.$$

But $Y_C = a + bX$; hence $\Sigma YY_C = \Sigma[Y(a + bX)] = \Sigma(aY + bXY)$

$$= a\Sigma Y + b\Sigma XY.$$

Now $a\Sigma Y + b\Sigma XY = \Sigma Y_C^2$ (equation 3).

Therefore $\Sigma y_S^2 = \Sigma Y^2 - 2\Sigma Y_C^2 + \Sigma Y_C^2$

$$= \Sigma Y^2 - \Sigma Y_C^2 ... .(6)$$

$$= \Sigma Y^2 - (a\Sigma Y + b\Sigma XY) .. ..(7)$$

To prove that $\sigma_y^2 = \sigma_{y_c}^2 + \sigma_{y_s}^2$.

This expression may be written:

$$\frac{\Sigma y^2}{N} = \frac{\Sigma y_c^2}{N} + \frac{\Sigma y_s^2}{N}.$$

Multiplying by N, we have

$$\Sigma y^2 = \Sigma y_c^2 + \Sigma y_s^2.$$

But we know that $\Sigma y^2 = \Sigma Y^2 - \overline{Y}\Sigma Y$, and it has been shown that $\Sigma y_c^2 = \Sigma Y_c^2 - \overline{Y}\Sigma Y$ (equation 4), and that $\Sigma y_s^2 = \Sigma Y^2 - \Sigma Y_C^2$ (equation 6).

Therefore, substituting, we have

$$\Sigma Y^2 - \overline{Y}\Sigma Y = \Sigma Y_C^2 - \overline{Y}\Sigma Y + \Sigma Y^2 - \Sigma Y_C^2$$
$$= \Sigma Y^2 - \overline{Y}\Sigma Y.$$

Thus $\Sigma y^2 = \Sigma y_C^2 + \Sigma y_S^2$, ... (8)

and $\sigma_y^2 = \sigma_{y_c}^2 + \sigma_{y_s}^2$ .. .. (9)

Section XXII-2

Derivation of Constants for Straight Line Equation when Origin is at $\overline{X}, \overline{Y}$

The normal equations for fitting a straight line by the method of least squares are

$$\Sigma Y = Na + b\Sigma X;$$
$$\Sigma XY = a\Sigma X + b\Sigma X^2.$$

If the origin be taken at $\overline{X}, \overline{Y}$ instead of 0,0 we have

$$\Sigma y = Na + b\Sigma x;$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2.$$

But $\Sigma y = 0$, and $\Sigma x = 0$.

Therefore $a = 0$, and $b = \dfrac{\Sigma xy}{\Sigma x^2}$.

The estimating equation becomes $y_C = bx$ instead of $Y_C = a + bX$.

Section XXII-3

Given that $r = \sqrt{\dfrac{\Sigma y_C^2}{\Sigma y^2}}$. To prove that $r = \dfrac{\Sigma xy}{N\sigma_x\sigma_y}$.

$$r^2 = \frac{\Sigma y_C^2}{\Sigma y^2}.$$

It follows from equation 5 of section XXII-1 that

$$\Sigma y_C^2 = b\Sigma xy$$
$$= \frac{\Sigma xy}{\Sigma x^2}\Sigma xy = \frac{(\Sigma xy)^2}{\Sigma x^2}.$$

Therefore

$$r^2 = \frac{(\Sigma xy)^2}{\Sigma x^2} \times \frac{1}{\Sigma y^2} = \frac{(\Sigma xy)^2}{\Sigma x^2 \Sigma y^2}.$$

But $\Sigma x^2 = N\sigma_x^2$, and $\Sigma y^2 = N\sigma_y^2$.

Hence

$$r^2 = \frac{(\Sigma xy)^2}{N^2 \sigma_x^2 \sigma_y^2}, \text{ and}$$

$$r = \frac{\Sigma xy}{N\sigma_x \sigma_y}.$$

Section XXII-4

To prove that $\dfrac{\Sigma xy}{N\sigma_x \sigma_y} = \dfrac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}}.$

$$\begin{aligned}
\Sigma xy &= \Sigma[(X - \overline{X})(Y - \overline{Y})] = \Sigma(XY - \overline{X}Y - X\overline{Y} + \overline{X}\,\overline{Y}), \\
&= \Sigma XY - \overline{X}\Sigma Y - \overline{Y}\Sigma X + N\overline{X}\,\overline{Y} \\
&= \Sigma XY - N\overline{X}\,\overline{Y} - N\overline{X}\,\overline{Y} + N\overline{X}\,\overline{Y} \\
&= \Sigma XY - N\overline{X}\,\overline{Y}.
\end{aligned}$$

$$\sigma_x = \sqrt{\frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2}, \text{ and } \sigma_y = \sqrt{\frac{\Sigma Y^2}{N} - \left(\frac{\Sigma Y}{N}\right)^2}.$$

Therefore

$$\begin{aligned}
\frac{\Sigma xy}{N\sigma_x \sigma_y} &= \frac{\Sigma XY - N\overline{X}\,\overline{Y}}{N\sqrt{\dfrac{\Sigma X^2}{N} - \left(\dfrac{\Sigma X}{N}\right)^2}\sqrt{\dfrac{\Sigma Y^2}{N} - \left(\dfrac{\Sigma Y}{N}\right)^2}} \\[2em]
&= \frac{N(\Sigma XY - N\overline{X}\,\overline{Y})}{\left[N\sqrt{\dfrac{\Sigma X^2}{N} - \left(\dfrac{\Sigma X}{N}\right)^2}\right]\left[N\sqrt{\dfrac{\Sigma Y^2}{N} - \left(\dfrac{\Sigma Y}{N}\right)^2}\right]} \\[2em]
&= \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}}.
\end{aligned}$$

Section XXII-5

Changing Units and Shifting Origin for a Straight Line Equation, Correlation of Grouped Data

The equation $d'_{Y_c} = .3561 - .4737\,d'_X$ is stated in terms of class-interval deviations from assumed means. Thus the value of Y at the assumed mean of X (44.5) is .3561 intervals above the assumed mean of Y (65).

To find the value of a when X (rather than d_X') is zero, we must therefore substitute -4.45 for d_X' in the equation. Thus

$$d_{Y_C}' = .3561 + (-4.45)(-.4737)$$
$$= 2.4641.$$

When X is zero, a is therefore 2 4641 intervals above the assumed mean of Y. It is thus 2.4641×10 physical units above 65, or 89.641.

In the original equation, b indicates how many *intervals* Y increases with an increase of one *interval* of X. To put b in terms of the original units, we must multiply b by the ratio of the Y interval to the X interval, in order to get b into terms of physical units. In this case, however, the class interval of each variable is 10, so that the multiplication does not change the value of b.

The equation then is $Y_C = 89.64 - .4737X$.

Section XXII-6

Derivation of Expression for Population Estimate of r

$$r^2 = 1 - \frac{\sigma_{y_s}^2}{\sigma_y^2}, \text{ and } \bar{r}^2 = 1 - \frac{\bar{\sigma}_{y_s}^2}{\bar{\sigma}_y^2}.$$

But $\bar{\sigma}_y^2 = \dfrac{\Sigma y^2}{N - 1} = \dfrac{N\sigma_y^2}{N - 1}$, since the y deviations are measured from the

mean, which is one constant; and $\bar{\sigma}_{y_s}^2 = \dfrac{\Sigma y_s^2}{N - m} = \dfrac{N\sigma_{y_s}^2}{N - m}$, since the y_s

deviations are measured from an estimating line containing m constants. Therefore

$$\bar{r}^2 = 1 - \frac{N\sigma_{y_s}^2 \div (N - m)}{N\sigma_y^2 \div (N - 1)}$$

$$= 1 - \frac{\sigma_{y_s}^2 (N - 1)}{\sigma_y^2 (N - m)}$$

$$= 1 - (1 - r^2)\frac{N - 1}{N - m}$$

$$= \frac{(N - m) - (1 - r^2)(N - 1)}{N - m}$$

$$= \frac{r^2(N - 1) - (m - 1)}{N - m}.$$

In the above form the expression may be used for linear or nonlinear

correlation, for multiple correlation, and for the correlation ratio. For simple linear correlation, $m = 2$; therefore we may write

$$\bar{r}^2 = \frac{r^2(N-1) - (2-1)}{N-2}$$

$$= \frac{r^2(N-1) - 1}{N-2}.$$

The derivation of the form used for the population estimate of a coefficient of partial correlation is shown in Section XXIV–4 of this Appendix.

Section XXIII–1

The flexibility of price, according to the equation $Y_C = 1,735,128X^{-2.149117}$ is -2.149117 throughout. With the other types of equations used in this chapter, the flexibility varies with quantity sold.

Flexibility of price is the percentage change in price (Y_C) associated with an infinitely small percentage change in quantity (X), or

$$\text{Flexibility} = \frac{dY_C \div Y_C}{dX \div X}.$$

This expression is more convenient to compute if written

$$\text{Flexibility} = \frac{dY_C}{dX} \times \frac{X}{Y_C}.$$

Therefore to compute the flexibility of price, differentiate with respect to X and multiply the derivative by the ratio of $\frac{X}{Y_C}$ at the point desired.

In the present instance we have

$$\text{Flexibility} = \left[(-2.149117)(1,735,128)X^{-3.149117}\right]\left[\frac{X}{1,735,128X^{-2.149117}}\right]$$

$$= -2.149117.$$

Section XXIII–2

The point of diminishing returns is the highest point in the curve: $Y_C = 890.324 + 78.264X + 20.324X^2 - 4.4649X^3$. At this point the slope is zero. The slope of a curve at any point may be found by taking the first derivative of the equation. The first derivative of the above equation is:

$$\frac{dY_C}{dX} = 78.264 + 40.648X - 13.3947X^2.$$

Setting $\dfrac{dY_c}{dX} = 0$, we have

$$78.264 + 40.648X - 13.3947X^2 = 0, \text{ and}$$

$$X = \frac{-40.648 \pm \sqrt{(40\ 648)^2 - 4(-13.3947)(78.264)}}{2(-13.3947)}$$

$$= 4.37128, \text{ or } -1.33669.$$

When the slope is zero, we have a maximum or a minimum point. In this case only positive values of X are of interest, and inspection of Chart 228 indicates that a maximum is reached when X is close to 4. Or, if the reader will compute Y_c values in the neighborhood of $X = -1.33669$ and $X = 4.37128$, he will discover that the former is a minimum and the latter a maximum.

When $X = 4.37128$,

$$Y_c = 890.324 + 78.264(4.37128) + 20.324(4.37128)^2 - 4.4649(4.37128)^3,$$
$$= 1,247.85.$$

The point of diminishing total returns is 4.37128 per cent nitrogen. At this point the estimated yield is 1,247.85 pounds.

The point of diminishing marginal returns is the point of inflection in the curve. It is the point where the change in the slope is zero. The change in the slope is the second derivative of the estimating equation. Thus

$$\frac{d^2Y_c}{dX^2} = 40.648 - 26.7894X.$$

Setting $\dfrac{d^2Y_c}{dX^2} = 0,$

$$40.648 - 26.7894X = 0, \text{ and}$$

$$X = 1.517317.$$

This is the point of diminishing marginal returns. At this point

$$Y_c = 890.324 + 78.264(1.517317) + 20.324(1.517317)^2 - 4.4649(1.517317)^3$$
$$= 1,040.27.$$

Section XXIII-3

Changing Units and Shifting Origin for a Second Degree Curve

The equation as stated is:

$$d'_{Y_c} = -.55290435 - .40268355d'_x + .058222561(d'_x)^2$$

Origin $\overline{X}_d$, $\overline{Y}_d$. Units: intervals of X and Y.

To obtain the equation in original units, but with origin at $\bar{X}_d$, $\bar{Y}_d$.

$$a = -.55290435 \times 25 = -13.82261;$$

$$b = -.40268355 \frac{25}{66.67} = -.1510063;$$

$$c = .058222561 \frac{25}{(66.67)^2} = .0003275019;$$

and the equation becomes

$$d_{Y_c} = -13.82261 - .1510063 d_X + .0003275019 (d_X)^2.$$
Origin: $X = 633.33$, $Y = 112\,5$.

To shift origin to $X = 0$, $Y = 0$, find the d_{Y_c} value when $X = -633.33$
Thus

$$d_{Y_c} = 13.82261 - .1510063(-633\,33) + .0003275019(633.33)^2$$
$$= 213.17936.$$
$$a = 213.17936 + 112.5 = 325.679.$$

But the slope of the line will also be different when $X = 0$.

To find the slope of the line: $Y_C = a + bX + cX^2$, differentiate with respect to X and substitute the desired value of X (in this case, -633.33).

$$\frac{dY_C}{dX} = b + 2cX = -.1510063 + 2\,(.0003275019)(-633.33)$$
$$= -.5658420.$$

It is not necessary to find a new value for c, however, as the change in the slope is the same throughout.

Therefore $Y_C = 325.679 - .5658420X + .0003275019X^2$.

Section XXIV-1

To prove that $a_{1.23} = \bar{X}_1 - b_{12.3}\bar{X}_2 - b_{13\,2}\bar{X}_3$.
Normal equation I is

$$\Sigma X_1 = Na_{1.23} + b_{12.3}\Sigma X_2 + b_{13.2}\Sigma X_3.$$

But $x = X - \bar{X}$, and $X = x + \bar{X}$. Therefore we may write for normal equation I:

$$\Sigma(x_1 + \bar{X}_1) = Na_{1.23} + b_{12\,3}\Sigma(x_2 + \bar{X}_2) + b_{13\,2}\Sigma(x_3 + \bar{X}_3), \text{ or}$$
$$\Sigma x_1 + N\bar{X}_1 = Na_{1.23} + b_{12.3}(\Sigma x_2 + N\bar{X}_2) + b_{13\,2}(\Sigma x_3 + N\bar{X}_3).$$

But $\Sigma x = 0$. Thus

$$N\bar{X}_1 = Na_{1.23} + b_{12.3}N\bar{X}_2 + b_{13\,2}N\bar{X}_3, \text{ and}$$
$$a_{1.23} = \bar{X}_1 - b_{12.3}\bar{X}_2 - b_{13.2}\bar{X}_3.$$

Section XXIV-2

Proof that

$$\left(\frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\ \sqrt{1 - r_{23}^2}}\right)^2 = \frac{\Sigma x_{C1\ 23}^2 - \Sigma x_{C1\ 3}^2}{\Sigma x_1^2 - \Sigma x_{C1.3}^2}$$

A demonstration for the other formulae of these types would proceed along similar lines.

If $r_{12.3} = \dfrac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\ \sqrt{1 - r_{23}^2}},$ ....(1)

$$r_{12\ 3}^2 = \frac{r_{12}^2 - 2r_{12}r_{13}r_{23} + r_{13}^2 r_{23}^2}{1 - r_{13}^2 - r_{23}^2 + r_{13}^2 r_{23}^2}.$$

But $r_{12}^2 = \dfrac{(\Sigma x_1 x_2)^2}{\Sigma x_1^2 \Sigma x_2^2}$, $r_{12} = \dfrac{\Sigma x_1 x_2}{\sqrt{\Sigma x_1^2 \Sigma x_2^2}}$, and similar formulae obtain for the

other r's.   Therefore:

$$r_{12\ 3}^2 = \frac{\dfrac{(\Sigma x_1 x_2)^2}{\Sigma x_1^2 \Sigma x_2^2} - 2\left[\dfrac{\Sigma x_1 x_2}{\sqrt{\Sigma x_1^2 \Sigma x_2^2}} \times \dfrac{\Sigma x_1 x_3}{\sqrt{\Sigma x_1^2 \Sigma x_3^2}} \times \dfrac{\Sigma x_2 x_3}{\sqrt{\Sigma x_2^2 \Sigma x_3^2}}\right] + \left[\dfrac{(\Sigma x_1 x_3)^2}{\Sigma x_1^2 \Sigma x_3^2} \times \dfrac{(\Sigma x_2 x_3)^2}{\Sigma x_2^2 \Sigma x_3^2}\right]}{1 - \dfrac{(\Sigma x_1 x_3)^2}{\Sigma x_1^2 \Sigma x_3^2} - \dfrac{(\Sigma x_2 x_3)^2}{\Sigma x_2^2 \Sigma x_3^2} + \left[\dfrac{(\Sigma x_1 x_3)^2}{\Sigma x_1^2 \Sigma x_3^2} \times \dfrac{(\Sigma x_2 x_3)^2}{\Sigma x_2^2 \Sigma x_3^2}\right]}$$

Multiplying numerator and denominator by $\Sigma x_1^2 \Sigma x_2^2 (\Sigma x_3^2)^2$ this simplifies to the following equation:

$$r_{12.3}^2 = \frac{(\Sigma x_3^2)^2 (\Sigma x_1 x_2)^2 - 2\Sigma x_3^2 \Sigma x_1 x_2 \Sigma x_1 x_3 \Sigma x_2 x_3 + (\Sigma x_1 x_3)^2 (\Sigma x_2 x_3)^2}{\Sigma x_1^2 \Sigma x_2^2 (\Sigma x_3^2)^2 - \Sigma x_2^2 \Sigma x_3^2 (\Sigma x_1 x_3)^2 - \Sigma x_1^2 \Sigma x_3^2 (\Sigma x_2 x_3)^2 + (\Sigma x_1 x_3)^2 (\Sigma x_2 x_3)^2} \quad (2)$$

$$r_{12\ 3}^2 = \frac{\Sigma x_{C1\ 23}^2 - \Sigma x_{C1\ 3}^2}{\Sigma x_1^2 - \Sigma x_{C1\ 3}^2} \quad ....(3)$$

But $\Sigma x_{C1.3}^2 = b_{13}\Sigma x_1 x_3 = \dfrac{\Sigma x_1 x_3}{\Sigma x_3^2} \Sigma x_1 x_3 = \dfrac{(\Sigma x_1 x_3)^2}{\Sigma x_3^2}.$

Also $\Sigma x_{C1\ 23}^2 = b_{12.3}\Sigma x_1 x_2 + b_{13.2}\Sigma x_1 x_3.$

Now, the normal equations for deriving $b_{12\ 3}$ and $b_{13.2}$ are:

II. $\Sigma x_1 x_2 = b_{12.3}\Sigma x_2^2 + b_{13.2}\Sigma x_2 x_3;$

III. $\Sigma x_1 x_3 = b_{12.3}\ \Sigma x_2 x_3 + b_{13\ 2}\Sigma x_3^2.$

In order to solve for $b_{13.2}$, we may multiply equation II by $\Sigma x_2 x_3$, and equation III by Σx_2^2, and subtract equation II from equation III.   Thus

II. $\Sigma x_1 x_2 \Sigma x_2 x_3 = b_{12.3}\Sigma x_2^2 \Sigma x_2 x_3 + b_{13.2}(\Sigma x_2 x_3)^2$

III. $\underline{\Sigma x_1 x_3 \Sigma x_2^2 = b_{12.3}\Sigma x_2^2 \Sigma x_2 x_3 + b_{13\ 2}\Sigma x_2^2 \Sigma x_3^2}$

$\Sigma x_1 x_3 \Sigma x_2^2 - \Sigma x_1 x_2 \Sigma x_2 x_3 = b_{13.2}\Sigma x_2^2 \Sigma x_3^2 - b_{13.2}(\Sigma x_2 x_3)^2$

$$b_{13.2} = \frac{\Sigma x_1 x_3 \Sigma x_2^2 - \Sigma x_1 x_2 \Sigma x_2 x_3}{\Sigma x_2^2 \Sigma x_3^2 - (\Sigma x_2 x_3)^2}.$$

In a similar fashion we may solve for $b_{12\ 3}$. This involves multiplying equation II by Σx_3^2 and equation III by $\Sigma x_2 x_3$. By such a process we find that

$$b_{12\ 3} = \frac{\Sigma x_1 x_3 \Sigma x_2 x_3 - \Sigma x_1 x_2 \Sigma x_3^2}{(\Sigma x_2 x_3)^2 - \Sigma x_2^2 \Sigma x_3^2}.$$

Substituting these expressions for $b_{13\ 2}$ and $b_{12.3}$ in the equation for $\Sigma x_{C1\ 23}^2$, we have

$$\Sigma x_{C1.23}^2 = \frac{\Sigma x_1 x_3 \Sigma x_2 x_3 - \Sigma x_1 x_2 \Sigma x_3^2}{(\Sigma x_2 x_3)^2 - \Sigma x_2^2 \Sigma x_3^2}\ \Sigma x_1 x_2 + \frac{\Sigma x_1 x_3 \Sigma x_2^2 - \Sigma x_1 x_2 \Sigma x_2 x_3}{\Sigma x_2^2 \Sigma x_3^2 - (\Sigma x_2 x_3)^2}\ \Sigma x_1 x_3.$$

This simplifies to

$$\Sigma x_{C1\ 23}^2 = \frac{(\Sigma x_1 x_3)^2 \Sigma x_2^2 + (\Sigma x_1 x_2)^2 \Sigma x_3^2 - 2\Sigma x_1 x_2 \Sigma x_1 x_3 \Sigma x_2 x_3}{\Sigma x_2^2 \Sigma x_3^2 - (\Sigma x_2 x_3)^2}.$$

Now substituting our expressions for $\Sigma x_{C1\ 23}^2$ and $\Sigma x_{C1\ 3}^2$ in formula (3) we have

$$r_{12.3}^2 = \frac{\dfrac{(\Sigma x_1 x_3)^2 \Sigma x_2^2 + (\Sigma x_1 x_2)^2 \Sigma x_3^2 - 2\Sigma x_1 x_2 \Sigma x_1 x_3 \Sigma x_2 x_3}{\Sigma x_2^2 \Sigma x_3^2 - (\Sigma x_2 x_3)^2} - \dfrac{(\Sigma x_1 x_3)^2}{\Sigma x_3^2}}{\Sigma x_1^2 - \dfrac{(\Sigma x_1 x_3)^2}{\Sigma x_3^2}}.$$

Expanding and simplifying, this expression becomes equation (2). Therefore

$$\left(\frac{r_{12} - r_{13} r_{23}}{\sqrt{1 - r_{13}^2}\ \sqrt{1 - r_{23}^2}}\right)^2 = \frac{\Sigma x_{C1\ 23}^2 - \Sigma x_{C1.3}^2}{\Sigma x_1^2 - \Sigma x_{C1.3}^2}.$$

Section XXIV-3

To show that $R_{1\ 234}^2 = 1 - [(1 - r_{14}^2)(1 - r_{13\ 4}^2)(1 - r_{12.34}^2)]$.
In the expression

$$\sigma_{S1\ 234}^2 = \frac{\Sigma x_1^2\ (1 - r_{14}^2)(1 - r_{13.4}^2)(1 - r_{12\ 34}^2)}{N},$$

$(1 - r_{14}^2)(1 - r_{13\ 4}^2)(1 - r_{12.34}^2)$ is the proportion of variation that has not been explained, that is

$$(1 - r_{14}^2)(1 - r_{13\ 4}^2)(1 - r_{12\ 34}^2) = \frac{\Sigma x_1^2 - \Sigma x_{C1.234}^2}{\Sigma x_1^2}.$$

This is demonstrated as follows:

$$\left(1 - r_{14}^2\right)\left(1 - r_{13\ 4}^2\right)\left(1 - r_{12.34}^2\right) = \left(1 - \frac{\Sigma x_{C1\ 4}^2}{\Sigma x_1^2}\right)\left(1 - \frac{\Sigma x_{C1.34}^2 - \Sigma x_{C1\ 4}^2}{\Sigma x_1^2 - \Sigma x_{C1.4}^2}\right)$$

$$\left(1 - \frac{\Sigma x_{C1\ 234}^2 - \Sigma x_{C1\ 34}^2}{\Sigma x_1^2 - \Sigma x_{C1\ 34}^2}\right)$$

$$= \left(\frac{\Sigma x_1^2 - \Sigma x_{C1\ 4}^2}{\Sigma x_1^2} \right) \left(\frac{\Sigma x_1^2 - \Sigma x_{C1\ 4}^2 - \Sigma x_{C1\ 34}^2 + \Sigma x_{C1\ 4}^2}{\Sigma x_1^2 - \Sigma x_{C1\ 4}^2} \right)$$

$$\left(\frac{\Sigma x_1^2 - \Sigma x_{C1\ 34}^2 - \Sigma x_{C1\ 234}^2 + \Sigma x_{C1\ 34}^2}{\Sigma x_1^2 - \Sigma x_{C1\ 34}^2} \right)$$

$$= \frac{(\Sigma x_1^2 - \Sigma x_{C1\ 4}^2)(\Sigma x_1^2 - \Sigma x_{C1\ 34}^2)(\Sigma x_1^2 - \Sigma x_{C1\ 234}^2)}{\Sigma x_1^2 (\Sigma x_1^2 - \Sigma x_{C1\ 4}^2)(\Sigma x_1^2 - \Sigma x_{C1\ 34}^2)}$$

$$= \frac{\Sigma x_1^2 - \Sigma x_{C1\ 234}^2}{\Sigma x_1^2}.$$

The expression $R_{1\ 234}^2 = 1 - [(1 - r_{14}^2)(1 - r_{13\ 4}^2)(1 - r_{12\ 34}^2)]$ is of course 1 minus the proportion of variation that is unexplained, or the proportion of variation that has been explained.

Section XXIV-4

To prove that $\bar{r}_{14\ 23}^2 = \dfrac{r_{14\ 23}^2(N - m + 1) - 1}{N - m}.$

$$1 - \bar{r}_{14.23}^2 = \frac{1 - \bar{R}_{1.234}^2}{1 - \bar{R}_{1\ 23}^2}, \text{ and}$$

$$\bar{r}_{14.23}^2 = \frac{(1 - \bar{R}_{1\ 23}^2) - (1 - \bar{R}_{1.234}^2)}{1 - \bar{R}_{1\ 23}^2}.$$

But $\bar{R}_{1\ 234}^2 = 1 - (1 - R_{1.234}^2)\dfrac{N - 1}{N - m}.$

Also $\bar{R}_{1\ 23}^2 = 1 - (1 - R_{1.23}^2)\dfrac{N - 1}{N - (m - 1)}.$ (In this expression, $m - 1$

is used instead of m, since $R_{1.23}^2$ involves one less constant in the estimating equation than does $R_{1\ 234}^2$.) Therefore

$$\bar{r}_{14.23}^2 = \frac{\left\{1 - \left[1 - \left(1 - R_{1\ 23}^2\right)\dfrac{N-1}{N-m+1}\right]\right\} - \left\{1 - \left[1 - \left(1 - R_{1\ 234}^2\right)\dfrac{N-1}{N-m}\right]\right\}}{1 - \left[1 - \left(1 - R_{1.23}^2\right)\dfrac{N-1}{N-m+1}\right]}$$

$$= \frac{\left(1 - R_{1.23}^2\right)\left(\dfrac{N-1}{N-m+1}\right) - \left(1 - R_{1\ 234}^2\right)\left(\dfrac{N-1}{N-m}\right)}{\left(1 - R_{1\ 23}^2\right)\left(\dfrac{N-1}{N-m+1}\right)}$$

$$= 1 \backsim \left(\frac{1 - R_{1\ 234}^2}{1 - R_{1.23}^2}\right)\left(\frac{N - m + 1}{N - m}\right)$$

$$= 1 - \left(1 - r_{14\ 23}^2\right)\left(\frac{N - m + 1}{N - m}\right)$$

$$= 1 - \frac{(N - m + 1) - r_{14\ 23}^2(N - m + 1)}{N - m}$$

$$= \frac{N - m - N + m - 1 + r_{14\ 23}^2(N - m + 1)}{N - m}$$

$$= \frac{r_{14\ 23}^2(N - m + 1) - 1}{N - m}.$$

APPENDIX C

Aids to Calculation

Adding machine. Figure 1 shows a standard type of adding machine. Its operation is very simple. Suppose one wishes to add 132, 356, and 1072. The procedure is as follows: (1) Press the total key and pull the
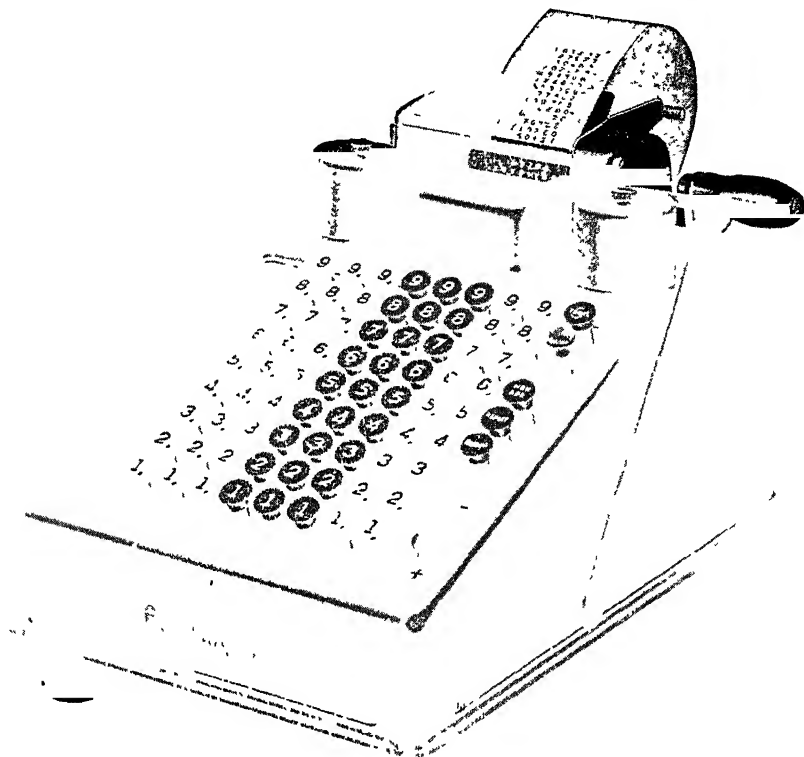
Figure 1. A Hand-Operated Burroughs Adding Machine.

lever in order to clear the machine. This is an important step, for if there is a number "in the machine," the final answer would be wrong by the amount of that number. If the machine has already been cleared, a star (without a number to the left of it) will now appear at the top of the adding machine slip. (2) In the last column of keys depress key num-

ber 2; in the adjoining row, key number 3; and in the third row from the
right, key 1.   (3)   Pull the large lever (or depress the activating key, if
machine is electric).   (4)   In similar fashion put 356 on the keyboard
and pull the lever.   (5)   Put 1072 on the keyboard and pull the lever.
(6)   Press the total key and pull the lever, thereby obtaining the total
and clearing the machine.   The adding machine slip appears thus:

```
              *
         132
         356
         1072
         1560 *
```

(The older style machines require that the lever be pulled again before
the total key is pressed.)   If desired, a subtotal of 132 and 356 may be
obtained by pressing the subtotal key after operation (4) above and pull-
ing the handle, with these results:

```
              *
         132
         356
         488 S
         1072
         1560 *
```

Most adding machines are now made with subtraction keys, which make
it possible to deduct values as desired.

   Calculating machine.   A machine that will both multiply and divide
(as well as add and subtract) is shown as Figure 2.

Figure 2.   A Hand-Operated Monroe Calculating Machine.

Addition and subtraction. To set the machine for addition or subtraction, release the repeat key. To add a number put it on the keyboard the same as with an adding machine and turn the handle forward (clockwise); to subtract it, turn the handle backward (counterclockwise). Turning the handle registers the result in the lower dial and clears the keyboard.

Multiplication. Multiplication may be thought of as repeated addition. Thus it would be possible to multiply 54 by 32 on the adding machine merely by putting 54 into the adding machine 32 times. Or it can be done as shown by the adding machine slip below:

```
        *
       54
       54
      540
      540
      540
     1728 *
```

The operation is performed in the same fashion by a calculating machine, but no printed slip is obtained. The process follows: (1) Set the machine for multiplication by pressing the repeat key. (2) Clear the machine, so that only zeros appear in the dials and no figures are depressed on the keyboard. (3) After the machine is cleared, 54 is put in the keyboard by depressing keys 5 and 4. (4) The large handle is turned forward twice. (5) The carriage is shifted once to the right and the handle turned forward three times. 54 now appears on the keyboard, 32 in the upper dial, and 1728 (the product) in the lower dial.

A short cut can be introduced when certain numbers are multiplied. For instance, to multiply 54 by 29, the easiest way is to multiply 54 by 30, and then subtract 54 once by turning the handle backward with the carriage at the extreme left.

Division. Division, which is merely repeated subtraction, is only slightly more complicated, though strictly analogous to long division by hand. If 1728 is to be divided by 32, the procedure is as follows: (1) Set the carriage several spaces to the right. (2) Set up 1728 in the keyboard and turn the handle once forward. 1728 will now appear in the lower dial and 1 in the upper. (3) Clear the 1 out of the upper dial by turning the small handle once forward (or by pressing the "clear" key and turning the large handle once backward). (4) Clear the keyboard. (5) Set up 32 on the keyboard so that the 3 in 32 will be in the same column of keys as the 7 in 1728. (Were the dividend 6728 the 3 would be placed under the 6.) (6) Turn the handle backward until the bell rings once, then turn it forward once, whereupon the bell will again ring. (7) Move the carriage one space to the left and repeat step 6. (8) Repeat step 7 as often as necessary. The answer appears in the upper dial.

As mentioned, the above method is equivalent to repeated subtraction. Thus:

$$
\begin{array}{rr}
& 1728 \\
& -\ 32 \\
\hline
\text{1st turn} & 1408 \\
& -\ 32 \\
\hline
\text{2nd turn} & 1088 \\
& -\ 32 \\
\hline
\text{3rd turn} & 768 \\
& -\ 32 \\
\hline
\text{4th turn} & 448 \\
& -\ 32 \\
\hline
\text{5th turn} & 128 \\
& -\ 32 \\
\hline
\text{1st turn} & 96 \\
& -\ 32 \\
\hline
\text{2nd turn} & 64 \\
& -\ 32 \\
\hline
\text{3rd turn} & 32 \\
& -\ 32 \\
\hline
\text{4th turn} & 0 \\
\end{array}
$$

5th turn — (Carriage is shifted at this point.)

The handle having been turned first five times and then four, the answer is 54.

Division by use of reciprocals. When a series of numbers is to be divided by the same number, the result can be obtained more easily by multiplying each of the numbers in turn by the reciprocal of the divisor. Thus:

$$1728 \div 32 = 1728 \times \frac{1}{32} = 1728 \times .03125 = 54.$$

The most frequent use of this method grows out of the need to express each of a series of numbers as a percentage of their total. The numbers 147, 265, and 376 total 788. In order to state each of these as a percentage of 788, we put the reciprocal of 788 $\left(\frac{1}{788} = .001269036\right)$ in the keyboard and multiply by 147. The result is .187, or 18.7 per cent. Then, *without clearing the keyboard* the upper dial is made to register 265, and the result is .336, or 33.6 per cent. Next, 376 is put in the upper dial, giving .477, or 47.7 per cent. (Frequently it will be convenient not to clear either the keyboard or the dials, and merely to change the upper dial successively. An entire problem is thus solved by this method without clearing the machine.)

Automatic electric machines. Electrically operated calculating machines with varying degrees of automatic control, which permit much more rapid calculation than do the manually operated types, are also available. Fig-

ures 3 and 4 illustrate two such machines that have automatic multiplication and division features. The principal of operation is essentially the same as with the manually operated machines, but the operation is easier.
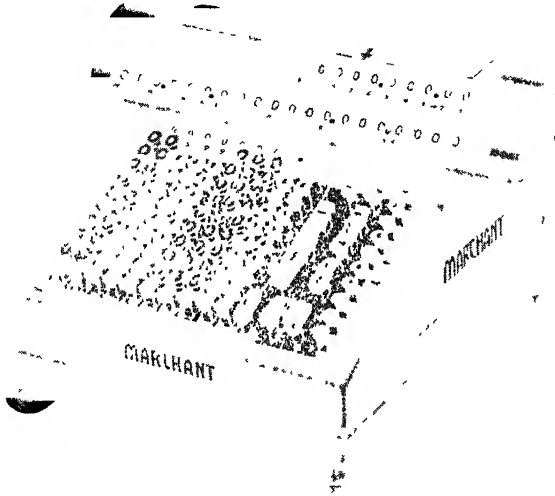
Figure 3.   An Automatic Electric Marchant Calculating Machine.
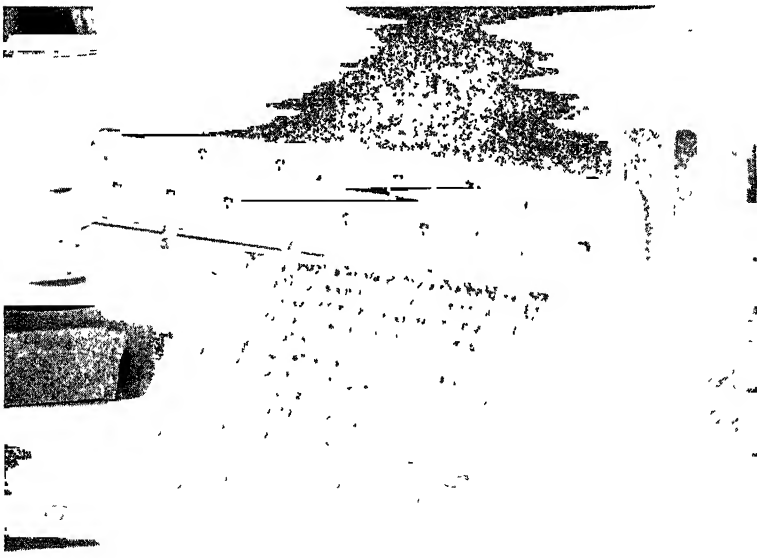
Figure 4.   An Automatic Electric Monroe Calculating Machine.

To divide, it is necessary only to put the dividend in the middle dial for the Marchant or the lower dial for the Monroe and the divisor on the keyboard, then actuate the machine by pressing the "divide" key (or lever). The quotient appears in the upper dial. In multiplying on the Marchant machine, one puts the multiplicand in the keyboard, and figures in the right-hand column corresponding to the multiplier are pressed in sequence. The product appears in the middle dial. Multiplication on the Monroe machine is similar to division, except that it is the "multiply" lever which is used and the product appears in the lower dial.

Marchant, Monroe, and other makes of machines which are less completely automatic than those pictured in Figures 3 and 4 but more fully equipped than that of Figure 2 may be bought.
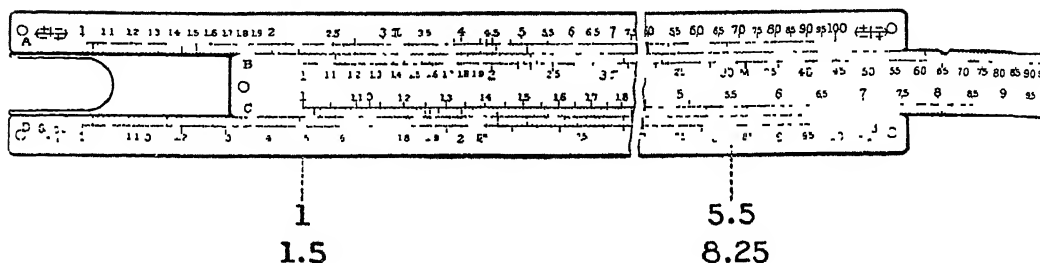
Figure 5.   A Slide Rule.

Slide rule. Unlike a calculating machine the slide rule is only approximately accurate, the accuracy depending on the size of the slide rule, the perfection of the materials and workmanship, and the skill of the user. The student should not expect to get more than three or four significant digits with a 10-inch slide rule. Also the slide rule does not automatically locate the decimal; this is ordinarily done very easily by inspection.

To multiply on a slide rule, set 1 on the slide (scale C) to the multiplicand on the rule (scale D), and opposite the multiplier on the slide (scale C) read the product on the rule (scale D). Figure 5 illustrates the multiplication of 1.5 by 5.5, the result being 8.25. The student will observe that a slide rule is merely one or more pairs of logarithmic scales. If he remembers that the principle of such scales is that equal distances represent equal proportions, he should recognize the principle by which the above computation was made. It is that 1 : 5.5 :: 1.5 : 8.25. (The setting of the slide rule in Figure 5 illustrates just as well, of course, that 15×55 = 825, or any other multiplication involving these digits.) Figure 6 illustrates the multiplication of 750 by 25, with a result of 18,750. The procedure is the same, except that it is necessary to set 10, instead of 1, on 75.

To divide, set the divisor on the slide (scale C) to the dividend on the rule (scale D), and read the quotient on the rule (scale D) opposite 1 or 10

on the slide (scale C). Figure 5 illustrates the computation $8.25 \div 5.5$ $= 1.5$, while Figure 6 illustrates $18,750 \div 25 = 75$.

To square, set the hair line of the cursor on scale D at the number to be squared and read the result from scale A at the point indicated by the hair line. An antithetical procedure is followed for extracting the square root. In Figure 6, the cursor is set to illustrate the squaring of 9, or the extraction of the square root of 81.

Tables. In this volume, use has been made of squares, sums of squares, square roots, reciprocals, and logarithms. Appendix O gives squares,
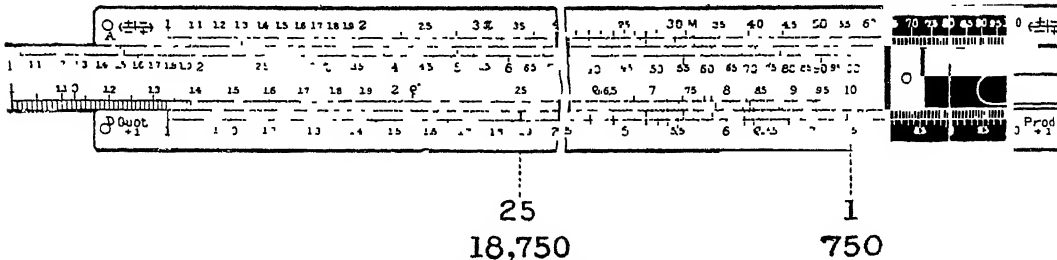
Figure 6. A Slide Rule.

square roots, and reciprocals for numbers from 1 to 1,000. Appendices M and N give the sums of powers of the first 100 natural numbers and of the first 100 odd natural numbers. A brief table of logarithms is shown as Appendix P.

More detailed tables of powers, roots, and reciprocals are given in *Barlow's Tables*, published by Spon and Chamberlain, New York.

A seven-place table of logarithms is given by James W. Glover in his *Tables of Applied Mathematics in Finance, Insurance, Statistics*, published by George Wahr, Ann Arbor, Michigan.

Various useful tables will also be found in R. A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural, and Medical Research*, published by Oliver and Boyd, Ltd., London.

APPENDIX D

Ordinates of the Normal Probability Curve

Erected at Distances $\frac{x}{\sigma}$ from the Mean, Expressed as Decimal Fractions of the Maximum Ordinate Y_0

The maximum ordinate is computed from the expression $Y_0 = \dfrac{Ni}{\sigma\sqrt{2\pi}} = \dfrac{Ni}{2.5066\sigma}$.

The values tabled below result from solving the expression $e^{\frac{-x^2}{2\sigma^2}}$.

The proportional height of an ordinate to be erected at any given value on the X axis can be read from the table by determining x (the deviation of the given value from the mean) and computing $\frac{x}{\sigma}$. Thus if $\overline{X} = \$25.00$, $\sigma = \$4.00$, $Y_0 = 1950$ and it is desired to ascertain the height of an ordinate to be erected at $\$23.00$; $x = \$2.00$ and $\dfrac{x}{\sigma} = \dfrac{\$2.00}{\$4.00} = .50$. From the table the ordinate is found to be .88250 of the maximum ordinate Y_0 or $.88250 \times 1950 = 1721$.

$\frac{x}{\sigma}$	0	1	2	3	4	5	6	7	8	9
0 0	1 00000	99995	99980	99955	.99920	99875	99820	99755	99685	99596
0 1	99501	99396	99283	99158	99025	.98881	98728	98565	98393	98211
0 2	98020	97819	97609	97390	.97161	96923	96676	96420	96156	95882
0.3	95600	95309	95010	94702	94387	94055	93723	93382	93024	92677
0 4	.92312	91939	91558	91169	.90774	90371	89961	89543	89119	88688
0 5	.88250	87805	87353	86896	.86432	85962	85488	85006	84519	84060
0 6	83527	83023	82514	82010	81481	80957	80429	.79896	79359	78817
0 7	.78270	77721	77167	76610	76048	75484	74916	.74342	73769	73193
0 8	.72615	72033	71448	70861	.70272	69681	69087	68493	.67896	67298
0 9	66689	66097	65494	64891	.64287	63683	63077	.62472	61865	61259
1 0	.60653	60047	59440	58834	58228	57623	57017	56414	55810	55209
1 1	.54607	54007	.53409	52812	52214	51620	51027	50437	.49848	49260
1 2	48675	.48092	47511	46933	46357	45783	45212	44644	.44078	43516
1 3	.42956	42399	.41845	41294	.40747	40202	39661	39123	.38569	38058
1 4	.37531	37007	36487	35971	35459	34950	34445	33944	33447	.32954
1 5	.32465	31980	.31500	.31023	30550	30082	29618	29158	.28702	.28251
1 6	.27804	27361	.26923	26489	.26059	.25634	.25213	24797	24385	.23978
1 7	23575	23176	22782	22392	22008	21627	21251	20879	20511	.20148
1 8	.19790	19436	.19086	18741	18400	18064	17732	.17404	.17081	.16762
1 9	.16448	16137	15831	.15530	15232	14939	14650	14364	14083	.13806
2.0	.13534	.13265	.13000	.12740	12483	.12230	11981	11737	11496	.11259
2.1	.11025	.10795	.10570	.10347	10129	09914	09702	.09495	09290	.09090
2.2	.08892	08698	08507	.08320	08136	07956	07778	07604	07433	07265
2 3	.07100	06939	.06780	06624	06471	06321	06174	.06029	05888	.05750
2.4	.05614	.05481	05350	05222	.05096	.04973	04852	04734	.04618	.04505
2.5	.04394	.04285	04179	.04074	.03972	.03873	03775	03680	03586	.03494
2 6	.03405	03317	03232	03148	.03066	.02986	.02908	.02831	02757	02684
2 7	02612	.02542	02474	02408	.02343	02280	02218	02157	.02098	.02040
2 8	.01984	01929	.01876	.01823	01772	.01723	01674	.01627	01581	01536
2 9	.01492	01449	01408	.01367	.01328	.01288	.01252	.01215	.01179	.01145
3.	01111	.00819	.00598	00432	.00309	.00219	00153	00106	00073	00050
4.	.00034	00022	00015	.00010	00006	.00004	00003	.00002	.00001	00001
5.	.00000									

NOTE: After $\frac{x}{\sigma} = 3.0$, ordinates are shown for steps of $.1\frac{x}{\sigma}$ instead of $.01\frac{x}{\sigma}$.

From Rugg's *Statistical Methods Applied to Education*, reprinted by arrangement with the publishers, Houghton Mifflin Company. A more detailed table of normal curve ordinates may be found in Karl Pearson, *Tables for Statisticians and Biometricians*, pp. 2–8, The University Press, Cambridge, England, 1914. The values shown in Pearson's table should be multiplied by $\sqrt{2\pi} = 2.5066$ to agree with those shown above.

APPENDIX E

Areas Under the Normal Probability Curve

From the Mean to Distances $\frac{x}{\sigma}$ from the Mean, Expressed as Decimal Fractions of the

Total Area 1.0000

The proportional part of the curve included between an ordinate erected at the mean and an ordinate erected at any given value on the X axis can be read from the table by determining x (the deviation of the given value from the mean) and computing $\frac{x}{\sigma}$.

Thus if $\overline{X} = \$25.00$, $\sigma = \$4.00$, and it is desired to ascertain the proportion of the area under the curve between ordinates erected at the mean and at $\$20.00$; $x = \$5.00$ and $\frac{x}{\sigma} = \frac{\$5.00}{\$4.00} = 1.25$. From the table it is found that .3944, or 39.44 per cent, of the entire area is included.

$\frac{x}{\sigma}$	00	01	.02	03	04	05	06	07	08	09
0 0	.0000	.0040	.0080	0120	0160	.0199	0239	0279	0319	.0359
0 1	0398	0438	0478	0517	0557	.0596	0636	.0675	.0714	.0753
0 2	0793	0832	0871	1910	0948	.0987	.1026	.1064	.1103	.1141
0 3	1179	1217	1255	1293	1331	.1368	.1406	.1443	.1480	.1517
0 4	1554	1591	1628	.1664	.1700	1736	.1772	.1808	1844	.1879
0 5	1915	1950	1985	.2019	2054	.2088	2123	.2157	2190	.2224
0 6	2257	.2291	2324	2357	.2389	.2422	2454	.2486	.2518	.2549
0.7	.2580	2612	2642	2673	.2704	2734	.2764	.2794	2823	.2852
0 8	2881	2910	2939	.2967	.2995	3023	3051	3078	3106	.3133
0 9	3159	3186	3212	3238	3264	3289	3315	3340	.3365	.3389
1 0	3413	3438	3461	3485	3508	3531	3554	3577	3599	.3621
1 1	3643	3665	.3686	.3708	3729	3749	.3770	3790	.3810	.3830
1 2	3849	3869	3888	.3907	3925	3944	3962	.3980	3997	.4015
1 3	4032	.4049	.4066	4082	4099	.4115	4131	4147	4162	.4177
1.4	4192	.4207	.4222	.4236	.4251	4265	4279	.4292	4306	4319
1.5	4332	4345	4357	4370	4382	4394	4406	.4418	.4429	.4441
1.6	4452	.4463	4474	4484	4495	4505	4515	4525	4535	4545
1.7	.4554	4564	4573	.4582	4591	4599	4608	.4616	4625	4633
1 8	.4641	4649	.4656	4664	4671	4678	.4686	4693	.4699	.4706
1 9	4713	4719	.4726	.4732	4738	4744	4750	4756	.4761	4767
2 0	.4772	4778	4783	.4788	4793	.4798	4803	4808	.4812	.4817
2.1	.4821	4826	4830	.4834	.4838	4842	4846	.4850	.4854	.4857
2 2	.4861	4864	4868	.4871	4875	4878	.4881	.4884	4887	4890
2 3	4893	.4896	.4898	.4901	4904	4906	4909	4911	.4913	.4916
2.4	4918	4920	4922	4925	4927	4929	4931	4932	4934	4936
2 5	4938	4940	.4941	4943	4945	.4946	4948	4949	.4951	4952
2 6	.4953	4955	4956	4957	4959	4960	4961	4962	4963	4964
2.7	4965	.4966	4967	4968	4969	4970	4971	.4972	.4973	4974
2.8	.4974	.4975	4976	4977	4977	4978	4979	4979	4980	.4981
2 9	4981	.4982	4982	4983	4984	4984	4985	4985	4986	4986
3 0	.49865	4987	4987	.4988	4988	4989	4989	4989	4990	.4990
3 1	.49903	4991	4991	4991	4992	.4992	4992	4992	4993	4993
3 2	.4993129									
3 3	4995166									
3 4	4996631									
3.5	4997674									
3 6	4998409									
3.7	4998922									
3.8	4999277									
3.9	.4999519									
4.0	4999683									
4.5	4999966									
5.0	4999997133									

From Rugg's *Statistical Methods Applied to Education*, reprinted by arrangement with the publishers, Houghton Mifflin Company.

Table of Values of t

For Given Degrees of Freedom (n) and at Specified Levels of Significance (P)

In the use of this table it is to be remembered that a level of significance refers to both tails of the distribution.   Thus, the .02 level ($P = .02$) includes .01 of the area of the curve in each tail   It is to be observed that this table is set up in a different form from the table of normal curve areas, Appendix E   The table of normal curve areas showed values of $\frac{x}{\sigma}$ in the margins and proportionate areas from $\overline{X}$ to $\frac{x}{\sigma}$ (one direction only) in the body   A tail of the normal distribution is obtained by subtracting this value from 5000   Doubling the resulting figure yields the level of significance The t table, on the other hand, shows n (degrees of freedom) in the stub, t in the body and P (the level of significance) in the caption.   The last row of the t table, for $N = \infty$ shows t values as obtained from the normal curve.

n	Level of Significance (P)												
	.9	.8	.7	.6	.5	.4	.3	.2	.1	.05	.02	.01	.001
1	.158	.325	.510	.727	1.000	1 376	1.963	3 078	6.314	12 706	31.821	63.657	636 619
2	.142	.289	445	617	.816	1 061	1 386	1.886	2 920	4 303	6 965	9 925	31 598
3	.137	.277	.424	.584	765	.978	1.250	1.638	2.353	3 182	4.541	5.841	12 941
4	134	271	.414	569	.741	.941	1 190	1 533	2 132	2.776	3 747	4 604	8 610
5	.132	267	.408	559	.727	.920	1.156	1.476	2.015	2.571	3 365	4.032	6 859
6	131	265	.404	.553	.718	.906	1.134	1.440	1.943	2.447	3.143	3 707	5 959
7	.130	263	402	549	.711	896	1.119	1.415	1 895	2 365	2 998	3 499	5 405
8	130	262	399	.546	706	889	1.108	1 397	1 860	2 306	2 896	3 355	5 041
9	.129	261	398	.543	703	.883	1 100	1 383	1.833	2 262	2.821	3 250	4 781
10	.129	.260	.397	.542	.700	.879	1.093	1 372	1 812	2 228	2.764	3.169	4 587
11	129	260	396	.540	.697	876	1 088	1.363	1 796	2 201	2 718	3.106	4 437
12	128	.259	.395	539	695	.873	1 083	1.356	1.782	2 179	2 681	3.055	4.318
13	128	.259	.394	538	694	.870	1 079	1 350	1 771	2 160	2.650	3.012	4.221
14	.128	.258	393	537	692	.868	1.076	1.345	1.761	2.145	2.624	2 977	4 140
15	.128	.258	393	.536	691	866	1.074	1 341	1 753	2 131	2.602	2.947	4 073
16	128	258	.392	535	690	.865	1.071	1.337	1 746	2.120	2 583	2.921	4 015
17	.128	257	.392	.534	689	.863	1.069	1 333	1 740	2.110	2 567	2.898	3 965
18	127	.257	.392	.534	688	.862	1.067	1 330	1.734	2 101	2.552	2 878	3 922
19	.127	257	.391	.533	.688	.861	1.066	1.328	1.729	2.093	2.539	2.861	3 883
20	.127	.257	391	.533	.687	.860	1.064	1.325	1.725	2.086	2.528	2 845	3.850
21	.127	.257	391	.532	.686	.859	1.063	1 323	1.721	2 080	2 518	2.831	3 819
22	.127	.256	.390	.532	.686	.858	1.061	1.321	1.717	2 074	2 508	2.819	3.792
23	.127	.256	.390	.532	685	859	1 060	1 310	1 714	2 069	2.500	2.807	3 767
24	.127	.256	390	.531	685	857	1 059	1 315	1 711	2 064	2 492	2 797	3 745
25	.127	.256	.390	.531	.684	.856	1.058	1.316	1.708	2.060	2.485	2.787	3 725
26	.127	.256	.390	.531	.684	.856	1 058	1 315	1.706	2.056	2.479	2.779	3 707
27	127	.256	.389	.531	.684	.855	1 057	1.314	1.703	2 052	2.473	2.771	3 690
28	127	256	.389	530	.683	.855	1 056	1 313	1 701	2.048	2 467	2.763	3.674
29	.127	.256	.389	.530	.683	.854	1 055	1 311	1.699	2.045	2.462	2.756	3 659
30	.127	.256	.389	.530	.683	.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	126	.255	388	.529	.681	.851	1 050	1.303	1.684	2 021	2 423	2.704	3.551
60	126	.254	.387	.527	.679	.848	1 046	1 296	1 671	2 000	2 390	2.660	3.460
120	126	251	386	526	.677	.845	1 041	1 289	1.658	1.980	2.358	2.617	3.373
∞	.126	253	385	321	.674	.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

APPENDIX G1

Values of z at the .05, .01, and .001 Points of the Distribution of z for Specified Values of n_1 and n_2

The term "point" is used to refer to one tail of a curve as a proportion of the entire area under consideration, in this case all positive values of z.

n_2	$n_1=1$ .05	.01	.001	$n_1=2$ .05	.01	.001	$n_1=3$ .05	.01	.001	$n_1=4$ .05	.01	.001	$n_1=5$ .05	.01	.001
1	2.5421	4.1535	6.4669	2.0479	4.2585	6.6612	2.6870	4.2974	6.6960	2.7071	4.3175	6.6201	2.7194	4.3297	6.6525
2	1.4592	2.2950	3.4531	1.4722	2.2976	3.4684	1.4765	2.2984	3.4635	1.4787	2.2988	3.4535	1.4800	2.2991	3.4535
3	1.1577	1.7649	2.1629	1.1284	1.7140	2.6003	1.1137	1.6915	2.4748	1.1051	1.6786	2.4603	1.1094	1.6703	2.4511
4	1.0212	1.5270	2.1629	.9690	1.4452	2.0574	.9429	1.4075	2.0143	.9272	1.3856	2.0892	.9108	1.3711	1.9728
5	.9441	1.3943	1.9255	.8777	1.2929	1.8602	.8441	1.2449	1.7513	.8236	1.2164	1.7184	.8097	1.1974	1.6964
6	.9013	1.3103	1.7849	.8188	1.1955	1.6479	.7798	1.1401	1.5828	.7558	1.1068	1.5433	.7394	1.0343	1.5177
7	.8660	1.2526	1.6874	.7777	1.1281	1.5384	.7347	1.0672	1.4603	.7080	1.0300	1.4221	.6896	1.0048	1.5387
8	.8443	1.2106	1.6177	.7475	1.0787	1.4687	.7014	1.0135	1.3809	.6725	.9734	1.3539	.6525	.9459	1.3008
9	.8401	1.1786	1.6646	.7242	1.0411	1.5982	.6757	.9724	1.3160	.6450	.9299	1.2053	.6238	.9006	1.2804
10	.8013	1.1535	1.6892	.7058	1.0114	1.3609	.6553	.9399	1.2650	.6232	.8954	1.2116	.6009	.8646	1.1748
11	.7880	1.1333	1.4900	.6900	.9874	1.3128	.6387	.9136	1.2238	.6055	.8674	1.1689	.5822	.8354	1.1897
12	.7788	1.1166	1.4687	.6786	.9677	1.2814	.6250	.8919	1.1900	.5907	.8443	1.1326	.5366	.8111	1.0026
13	.7703	1.1027	1.4400	.6682	.9511	1.2553	.6134	.8737	1.1616	.5783	.8248	1.1026	.5535	.7907	1.0614
14	.7640	1.0909	1.4208	.6594	.9370	1.2532	.6036	.8581	1.1376	.5677	.8082	1.0772	.5423	.7732	1.0348
15	.7548	1.0807	1.4043	.6518	.9249	1.2141	.5950	.8448	1.1169	.5585	.7939	1.0553	.5326	.7582	1.0119
16	.7514	1.0719	1.8000	.6451	.9144	1.1976	.5876	.8331	1.0989	.5505	.7814	1.0363	.5241	.7450	.9920
17	.7466	1.0641	1.3775	.6393	.9051	1.1832	.5811	.8229	1.0832	.5434	.7705	1.0195	.5166	.7335	.9745
18	.7424	1.0572	1.3665	.6341	.8970	1.1704	.5753	.8138	1.0693	.5371	.7607	1.0047	.5099	.7232	.9590
19	.7386	1.0511	1.3567	.6295	.8897	1.1591	.5701	.8057	1.0569	.5315	.7521	.9915	.5040	.7140	.9448
20	.7352	1.0457	1.3480	.6254	.8831	1.1489	.5654	.7985	1.0458	.5265	.7443	.9798	.4986	.7058	.9329
21	.7322	1.0408	1.3401	.6216	.8772	1.1398	.5612	.7920	1.0358	.5219	.7372	.9691	.4938	.6984	.9217
22	.7294	1.0363	1.3329	.6182	.8719	1.1315	.5574	.7860	1.0268	.5178	.7309	.9595	.4894	.6916	.9116
23	.7269	1.0322	1.3264	.6151	.8670	1.1240	.5540	.7806	1.0186	.5140	.7251	.9607	.4854	.6855	.9024
24	.7246	1.0285	1.3205	.6123	.8626	1.1171	.5508	.7757	1.0111	.5106	.7197	.9427	.4817	.6799	.8939
25	.7225	1.0251	1.3151	.6097	.8585	1.1108	.5478	.7712	1.0041	.5074	.7148	.9354	.4783	.6747	.8862
26	.7205	1.0220	1.3101	.6073	.8548	1.1050	.5451	.7670	.9978	.5045	.7103	.9286	.4752	.6699	.8791
27	.7187	1.0191	1.3055	.6051	.8513	1.0997	.5427	.7631	.9920	.5017	.7062	.9223	.4723	.6655	.8735
28	.7171	1.0164	1.3013	.6030	.8481	1.0947	.5403	.7595	.9866	.4992	.7023	.9165	.4696	.6614	.8664
29	.7155	1.0139	1.2973	.6011	.8451	1.0903	.5382	.7562	.9815	.4969	.6987	.9112	.4671	.6576	.8607
30	.7141	1.0116	1.2936	.5994	.8423	1.0859	.5362	.7531	.9768	.4947	.6954	.9061	.4648	.6540	.8554
60	.6933	.9784	1.2415	.5738	.8025	1.0248	.5073	.7086	.9100	.4632	.6472	.8345	.4311	.6028	.7798
∞	.6729	.9462	1.1910	.5486	.7636	.9663	.4787	.6651	.8453	.4319	.5999	.7648	.3974	.5522	.7069

876

n_2	$n_1=6$			$n_1=8$			$n_1=12$			$n_1=24$			$n_1=\infty$		
	.05	.01	.001	.05	.01	.001	.05	.01	.001	.05	.01	.001	.05	.01	.001
1	2.7276	4.3379	6.6405	2.7380	4.3482	6.6508	2.7484	4.3585	6.6611	2.7588	4.3689	6.6715	2.7693	4.3794	6.6819
2	1.4808	2.2992	3.4535	1.4819	2.2994	3.4535	1.4830	2.2997	3.4536	1.4840	2.2999	3.4536	1.4851	2.3001	3.4536
3	1.0953	1.6645	2.4446	1.0899	1.6569	2.4361	1.0842	1.6489	2.4272	1.0781	1.6404	2.4179	1.0716	1.6314	2.4081
4	.9093	1.3609	1.9612	.8993	1.3473	1.9459	.8885	1.3327	1.9294	.8767	1.3170	1.9118	.8639	1.3000	1.8927
5	.7997	1.1838	1.6808	.7862	1.1656	1.6596	.7714	1.1457	1.6370	.7550	1.1239	1.6123	.7368	1.0997	1.5845
6	.7274	1.0680	1.4986	.7112	1.0460	1.4730	.6931	1.0218	1.4449	.6729	.9948	1.4134	.6499	.9643	1.3783
7	.6761	.9864	1.3711	.6576	.9614	1.3417	.6369	.9335	1.3090	.6134	.9020	1.2721	.5862	.8658	1.2296
8	.6378	.9259	1.2770	.6175	.8983	1.2443	.5945	.8673	1.2077	.5682	.8319	1.1662	.5371	.7904	1.1169
9	.6080	.8791	1.2047	.5862	.8494	1.1694	.5613	.8157	1.1293	.5324	.7769	1.0830	.4979	.7305	1.0279
10	.5843	.8419	1.1475	.5611	.8104	1.1098	.5346	.7744	1.0668	.5035	.7324	1.0165	.4657	.6816	.9557
11	.5648	.8116	1.1018	.5406	.7785	1.0614	.5126	.7405	1.0157	.4795	.6958	.9619	.4387	.6408	.8967
12	.5487	.7864	1.0628	.5234	.7520	1.0213	.4941	.7122	.9733	.4592	.6649	.9162	.4156	.6061	.8460
13	.5350	.7652	1.0306	.5089	.7295	.9875	.4785	.6882	.9374	.4419	.6386	.8774	.3957	.5761	.8014
14	.5233	.7471	1.0031	.4964	.7103	.9586	.4649	.6675	.9066	.4269	.6159	.8439	.3782	.5500	.7635
15	.5131	.7314	.9795	.4855	.6937	.9336	.4532	.6496	.8800	.4138	.5961	.8147	.3628	.5269	.7301
16	.5042	.7177	.9588	.4760	.6791	.9119	.4428	.6339	.8567	.4022	.5786	.7891	.3490	.5064	.7005
17	.4964	.7057	.9407	.4676	.6663	.8927	.4337	.6199	.8361	.3919	.5630	.7664	.3366	.4875	.6740
18	.4894	.6950	.9246	.4602	.6549	.8757	.4255	.6075	.8178	.3827	.5491	.7462	.3253	.4712	.6502
19	.4832	.6854	.9103	.4535	.6447	.8605	.4182	.5964	.8014	.3743	.5366	.7277	.3151	.4560	.6285
20	.4776	.6768	.8974	.4474	.6355	.8469	.4116	.5864	.7867	.3668	.5253	.7115	.3057	.4421	.6086
21	.4725	.6690	.8858	.4420	.6272	.8346	.4055	.5773	.7735	.3599	.5150	.6964	.2971	.4294	.6004
22	.4679	.6620	.8753	.4370	.6196	.8234	.4001	.5691	.7612	.3536	.5056	.6828	.2892	.4176	.5758
23	.4636	.6555	.8657	.4325	.6127	.8132	.3950	.5615	.7501	.3478	.4969	.6704	.2818	.4068	.5583
24	.4598	.6496	.8569	.4283	.6064	.8038	.3904	.5545	.7400	.3425	.4890	.6589	.2749	.3967	.5440
25	.4562	.6442	.8489	.4244	.6006	.7953	.3862	.5481	.7306	.3376	.4816	.6483	.2685	.3872	.5307
26	.4529	.6392	.8415	.4209	.5952	.7873	.3823	.5422	.7220	.3330	.4748	.6385	.2625	.3784	.5183
27	.4499	.6346	.8346	.4176	.5902	.7800	.3786	.5367	.7140	.3287	.4685	.6294	.2562	.3701	.5066
28	.4471	.6303	.8282	.4146	.5856	.7733	.3752	.5316	.7066	.3248	.4626	.6209	.2516	.3624	.4967
29	.4444	.6263	.8225	.4117	.5813	.7679	.3720	.5269	.6997	.3211	.4570	.6129	.2466	.3550	.4863
30	.4420	.6226	.8168	.4090	.5773	.7610	.3691	.5224	.6932	.3176	.4519	.6056	.2419	.3481	.4756
60	.4064	.5687	.7877	.3702	.5189	.6760	.3255	.4574	.5992	.2654	.3746	.4955	.1644	.2352	.3198
∞	.3706	.5152	.6599	.3309	.4604	.6017	.2804	.3908	.5044	.2085	.2913	.3786	0	0	0

Based on tables in R. A. Fisher, *Statistical Methods for Research Workers*, by permission of Professor Fisher and Oliver and Boyd, Ltd   Values of z at the .20 point are shown in R. A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural and Medical Research*, p. 28, Oliver and Boyd, Ltd, Edinburgh, 1938.

In the above table the values of z at the .001 point for $n_1 = 1$, $n_2 = 1$ and for $n_1 = 12$, $n_2 = 2$ have been corrected to agree with those given in Fisher and Yates, *op cit.*, p. 34.

APPENDIX G 2

Values of F at the .05, .01, and .001 Points of the Distribution of F for Specified Values of n_1 and n_2

n_2	$n_1=1$			$n_1=2$			$n_1=3$			$n_1=4$			$n_1=5$		
	.05	.01	.001	.05	.01	.001	.05	.01	.001	.05	.01	.001	.05	.01	.001
1	161.45	4,052.1	405,300.	199.50	4,999.0	500,000	215.72	5,403.5	536,700.	224.57	5,625.1	562,500.	230.17	5,764.1	576,400.
2	18.512	98.495	998.4	18.999	99.008	999.0	19.163	99.167	999.2	19.248	99.246	999.2	19.298	99.305	999.2
3	10.129	34.117	167.46	9.552	30.815	148.5	9.276	29.459	141.11	9.118	28.709	137.08	9.014	28.236	134.58
4	7.710	21.200	74.126	6.945	18.001	61.238	6.591	16.693	56.181	6.388	15.978	53.430	6.257	15.521	51.706
5	6.607	16.258	47.039	5.786	13.274	36.612	5.410	12.059	33.201	5.192	11.391	31.087	5.050	10.966	29.748
6	5.987	13.744	35.509	5.143	10.924	26.998	4.756	9.779	23.702	4.534	9.149	21.903	4.388	8.746	20.809
7	5.591	12.246	29.218	4.737	9.546	21.688	4.347	8.452	18.772	4.121	7.846	17.188	3.972	7.460	16.206
8	5.317	11.259	25.416	4.450	8.649	18.493	4.067	7.591	15.888	3.838	7.006	14.388	3.688	6.631	13.485
9	5.117	10.561	22.855	4.256	8.022	16.385	3.863	6.992	13.901	3.633	6.423	12.561	3.482	6.057	11.714
10	4.965	10.044	21.030	4.103	7.560	14.906	3.708	6.552	12.553	3.478	5.994	11.282	3.326	5.636	10.481
11	4.844	9.647	19.687	3.982	7.205	13.813	3.587	6.217	11.560	3.357	5.668	10.346	3.204	5.317	9.577
12	4.747	9.330	18.641	3.885	6.927	12.972	3.490	5.953	10.805	3.259	5.412	9.633	3.106	5.064	8.899
13	4.667	9.074	17.814	3.805	6.701	12.312	3.410	5.740	10.208	3.179	5.205	9.072	3.025	4.862	8.354
14	4.600	8.862	17.143	3.739	6.514	11.780	3.344	5.563	9.730	3.112	5.035	8.625	2.958	4.695	7.922
15	4.543	8.683	16.586	3.683	6.359	11.338	3.287	5.417	9.335	3.056	4.893	8.253	2.901	4.556	7.567
16	4.494	8.532	16.119	3.634	6.227	10.970	3.239	5.292	9.005	3.007	4.772	7.944	2.853	4.437	7.272
17	4.451	8.400	15.721	3.592	6.112	10.659	3.197	5.185	8.727	2.965	4.669	7.683	2.810	4.336	7.022
18	4.414	8.285	15.379	3.555	6.013	10.389	3.160	5.092	8.487	2.928	4.579	7.459	2.773	4.248	6.807
19	4.381	8.184	15.080	3.522	5.926	10.157	3.127	5.010	8.280	2.895	4.501	7.264	2.740	4.170	6.609
20	4.351	8.096	14.820	3.493	5.849	9.953	3.098	4.938	8.098	2.866	4.431	7.102	2.711	4.103	6.461
21	4.325	8.017	14.588	3.467	5.780	9.773	3.072	4.875	7.937	2.840	4.368	6.946	2.685	4.042	6.318
22	4.301	7.944	14.379	3.443	5.719	9.612	3.049	4.816	7.796	2.817	4.314	6.814	2.661	3.988	6.192
23	4.279	7.881	14.194	3.422	5.663	9.469	3.028	4.765	7.669	2.795	4.264	6.695	2.640	3.939	6.079
24	4.260	7.823	14.027	3.403	5.614	9.339	3.009	4.718	7.555	2.777	4.218	6.589	2.621	3.895	5.976
25	4.242	7.770	13.875	3.385	5.568	9.223	2.991	4.676	7.450	2.759	4.177	6.493	2.603	3.855	5.885
26	4.225	7.722	13.738	3.369	5.527	9.116	2.975	4.637	7.356	2.743	4.140	6.406	2.587	3.818	5.803
27	4.210	7.677	13.612	3.354	5.488	9.020	2.961	4.601	7.272	2.727	4.106	6.326	2.572	3.785	5.726
28	4.196	7.636	13.498	3.340	5.453	8.930	2.947	4.568	7.194	2.714	4.074	6.253	2.558	3.754	5.656
29	4.183	7.597	13.391	3.328	5.421	8.852	2.934	4.538	7.121	2.702	4.045	6.187	2.545	3.726	5.593
30	4.171	7.563	13.293	3.316	5.390	8.774	2.922	4.510	7.054	2.690	4.018	6.124	2.534	3.699	5.535
60	4.001	7.077	11.972	3.151	4.978	7.765	2.758	4.126	6.172	2.525	3.649	5.307	2.368	3.339	4.757
∞	3.841	6.635	10.826	2.996	4.605	6.908	2.605	3.782	5.423	2.372	3.320	4.616	2.214	3.017	4.103

n_2	$n_1 = 6$			$n_1 = 8$			$n_1 = 12$			$n_1 = 24$			$n_1 = \infty$		
	.05	.01	.001	.05	.01	.001	.05	.01	.001	.05	.01	.001	.05	.01	.001
1	233 97	5,859 4	585,900	238 89	5,981.4	598,100	243 91	6,105.8	610,600	249 04	6,234 2	623,400	254 32	6,366 5	636,500
2	19 329	99.325	999 2	19 371	99.365	999 4	19 414	99.425	999 4	19 453	99.464	999 4	19 496	99.504	999 4
3	8 941	27.910	132.84	8 844	27.489	130 6	8 744	27 053	128 80	8 638	26 597	125 94	8 527	26.122	123 49
4	6 164	15.208	50 520	6 041	14.800	48 908	5 912	14 374	47 407	5 774	13 930	45 768	5 628	13.464	44 052
5	4 950	10.672	28 835	4 818	10.266	27 638	4 678	9 888	26 416	4 527	9.467	25 143	4 365	9.019	23 783
6	4 284	8.465	20.029	4 147	8.101	19.029	4 000	7.718	17 989	3 841	7.313	16 891	3 669	6.880	15 746
7	3 866	7.191	15 621	3 725	6.840	14 634	3 574	6.467	13 708	3 410	6.074	12 733	3 230	5.650	11 695
8	3 580	6.371	12 858	3 438	6.029	12 044	3 284	5.667	11 194	3 116	5.279	10 302	2 928	4.859	9 335
9	3 374	5.802	11 127	3 230	5.467	10 369	3 073	5.111	9 570	2 900	4.730	8 723	2 707	4.311	7 813
10	3 217	5.386	9 924	3 072	5.057	9 204	2 913	4.706	8 445	2 737	4.327	7 637	2 538	3.909	6 762
11	3 094	5.069	9 047	2 948	4.745	8 354	2 788	4.397	7 625	2 609	4 021	6 847	2 405	3 602	5 998
12	2 999	4.820	8 378	2 848	4.500	7 711	2 686	4.156	7 005	2 505	3 780	6 248	2 296	3 361	5 419
13	2 915	4.620	7 865	2 767	4.302	7 206	2 604	3.961	6 519	2 420	3 586	5 782	2 207	3 165	4 967
14	2 848	4.456	7 485	2 699	4.140	6 802	2 534	3.800	6 130	2 349	3 427	5 408	2 131	3 005	4 604
15	2 790	4.318	7 092	2 641	4.004	6 470	2 475	3.668	5 812	2 288	3 294	5 101	2 066	2 869	4 307
16	2 741	4.201	6 804	2 591	3.889	6 195	2 424	3.553	5 648	2 235	3.181	4 846	2 010	2 753	4 059
17	2 699	4.102	6 663	2 548	3.791	5 962	2 381	3.455	5 524	2 190	3.083	4 631	1 961	2 653	3 860
18	2 661	4.015	6 365	2 510	3.706	5 763	2 342	3.370	5 192	2 150	2.999	4 448	1 917	2 566	3 671
19	2 629	3.939	6 176	2 477	3.631	5 590	2 308	3.296	4 967	2 114	2.925	4 286	1 878	2 489	3 615
20	2 599	3.871	6 018	2 447	3.565	5 440	2 278	3.231	4 823	2 083	2.859	4 150	1 843	2 421	3 378
21	2 573	3.811	5 880	2 421	3.506	5 308	2 250	3.173	4 697	2 054	2.801	4 026	1 812	2.360	3 257
22	2 549	3.759	5 758	2 397	3.453	5 190	2 226	3.121	4 583	2 028	2.749	3 918	1 783	2.305	3 151
23	2 528	3.710	5 648	2 375	3.406	5 086	2 203	3.074	4 482	2 005	2.702	3 822	1 757	2.256	3 064
24	2 508	3.666	5 550	2 355	3.363	4 991	2 183	3.031	4 393	1 984	2.659	3 735	1 733	2.210	2 968
25	2 490	3.627	5 462	2 337	3.324	4 907	2 165	2.993	4 311	1 965	2.620	3 657	1 711	2.169	2 890
26	2 474	3.591	5 382	2 321	3.288	4 829	2 148	2.958	4 238	1 947	2.585	3 586	1 691	2 132	2 820
27	2 459	3.558	5 308	2 305	3.256	4 759	2 132	2.925	4 170	1 930	2.551	3 521	1 672	2 096	2 754
28	2 445	3.528	5 240	2 292	3.226	4 694	2 118	2.896	4 109	1 915	2.522	3 462	1 654	2 064	2 695
29	2 432	3.499	5 179	2 278	3.198	4 645	2 104	2.869	4 053	1 901	2.494	3 407	1 638	2 034	2 640
30	2 421	3.474	5 122	2 266	3.173	4 581	2 092	2.843	4 000	1 887	2.469	3 358	1 622	2 006	2 589
60	2 254	3.119	4 373	2 097	2.823	3 865	1 918	2.496	3 315	1 700	2 115	2 694	1 389	1.601	1 896
∞	2 099	2.802	3 743	1 938	2.511	3 265	1 752	2.182	2 742	1 517	1.791	2 132	1 000	1.000	1 000

Values of L at the .05 and .01 Levels of Significance for the Distribution of L for Specified Values of N and k, when $N_1 = N_2 = \cdots = N_k = N$

If L has been computed from samples of varying size, take $N = \dfrac{N_1 + N_2 + \cdots + N_k}{k}$, provided that no sample N is less than 15 or 20.

k	$N=3$ .05	.01	$N=4$ .05	.01	$N=5$ .05	.01	$N=6$ .05	.01	$N=7$ .05	.01	$N=8$ .05	01	$N=9$ .05	.01
2	312	.141	478	.284	585	398	656	.485	708	.551	745	603	775	.645
3	304	.162	470	.314	576	.429	648	.514	700	.578	739	628	769	.667
4	315	.188	480	.345	585	.459	656	.542	707	604	744	.652	774	689
5	328	.210	491	.370	595	.484	665	.565	714	.624	751	.670	780	.706
6	339	230	502	.391	604	.504	673	.583	721	.641	757	685	785	720
7	350	.246	512	.409	612	.520	680	.597	727	.654	763	697	790	.730
8	359	.260	520	424	620	534	686	610	733	.665	768	707	795	.740
9	367	.273	527	.437	626	545	691	.620	.738	.674	.772	715	798	.747
10	374	284	534	.448	631	.555	696	629	742	.682	776	.722	802	.753
12	387	.303	545	467	.641	572	704	.644	749	.696	782	734	807	.764
14	397	318	554	.481	649	.585	711	.655	755	.706	787	744	812	.773
16	405	331	561	493	655	.596	716	.665	759	.714	791	.751	816	.779
18	412	342	567	.504	660	.605	721	.672	763	.721	795	.756	819	.784
20	418	.352	573	.512	665	.613	725	.679	767	.727	798	.761	.822	.788
22	424	360	577	.520	669	.619	728	.684	770	.732	800	.765	824	.792
24	428	367	581	526	672	.624	731	688	772	.736	802	.768	826	.795
26	433	.373	585	532	675	.629	734	.693	775	.740	805	.772	828	.798
28	437	.379	589	.537	678	.634	736	697	777	.744	807	776	829	.802
30	441	.386	592	543	681	.639	739	.703	.779	.748	809	.781	831	.806

k	$N=10$ .05	.01	$N=12$ .05	.01	$N=15$ .05	.01	$N=20$ .05	.01	$N=30$ .05	.01	$N=60$ .05	.01	$N=\infty$ .05	.01
2	798	.678	833	730	868	.783	902	.836	935	.890	968	945	1.000	1.000
3	.792	699	.828	.748	863	.798	.898	.848	933	.898	967	.949	1 000	1.000
4	797	719	832	.765	866	.812	900	.859	934	.906	967	.953	1 000	1.000
5	.802	.735	836	779	.870	.823	.903	.867	936	.911	968	.956	1 000	1.000
6	808	.748	.841	.789	873	.832	906	.874	938	.916	969	.958	1 000	1.000
7	812	.757	844	.798	876	.839	908	.879	.939	.920	970	.960	1 000	1 000
8	816	.766	848	.805	.879	.844	910	.884	941	.923	971	.962	1 000	1.000
9	819	.773	851	811	881	.849	912	887	942	.925	971	.963	1 000	1.000
10	822	.779	853	.816	883	.853	913	.890	.943	.927	972	.964	1 000	1 000
12	828	.789	857	.824	887	.860	916	.896	944	.931	.973	.966	1 000	1.000
14	832	.796	861	.831	890	.865	918	.900	946	.933	973	.967	1 000	1.000
16	.835	.802	863	.836	892	870	920	903	947	.936	974	968	1 000	1 000
18	838	807	.866	.840	894	.873	921	.905	948	.937	974	.969	1 000	1.000
20	840	.811	868	.844	896	.876	922	.908	.949	.939	975	.970	1 000	1.000
22	843	.814	870	.847	897	.878	924	.909	950	.940	975	.970	1 000	1.000
24	844	.817	872	.850	898	.880	924	911	950	.941	975	.971	1.000	1.000
26	846	.820	873	.852	.809	.882	925	.912	.951	.942	.976	.971	1.000	1.000
28	848	.823	874	.854	900	.884	926	.914	951	.943	.976	.972	1 000	1.000
30	849	.827	876	.856	901	.886	.927	.915	.952	.944	.976	.972	1 000	1.000

APPENDIX I

Values of χ^2

For Given Degrees of Freedom (n) and for Specified Values of P

n	.99	.98	95	90	.80	70	50	30	.20	10	05	02	01	001.
1	000157	.000628	.00393	.0158	.0642	.148	.455	1 074	1 642	2.706	3 841	5 412	6 635	10 827
2	.0201	.0404	.103	.211	.446	.713	1 386	2 408	3 219	4 605				
3	115	.185	.352	.584	1.424	2.366	3 665	4 642	6.251					
4	.297	.429	711	1 064	1 649	2 195	3 357	4 878	5 989	7 779				
5	.554	.752	1 145	1 610	2 343	3 000	4 351	6 064	7 289	9 236	11 070	13 388	15 086	20 517
6	.872	1 134	1 635	2.204	3 070	3 828	5 348	7 231						
7	1 239	1 564	2 167	2 833	3.822	4 671	6 346	8 383						
8	1 646	2 032	2 733	3 490	4.594	5 527	7 344	9 524						
9	2 088	2 532	3 325	4 168	5 380	6.393	8.343	10 656						
10	2.558	3 059	3.940	4.865	6.179	7 267	9 342	11781						
11	3 053	3 609	4.575	5.578	6.989	8.148	10 341	12 899	14 631	17.275	19 675	22 618	24 725	31 264
12	3 571	4 178	5 226	6.304	7 807	9 034	11 340	14 011	15 812	18 549	21 026	24 054	26 217	32 909
13	4.107	4 765	5 892	7 042	8 634	9 926	12 340	15 119	16 985	19 812	22 362	25 472	27 688	34 528
14	4.660	5 368	6 571	7 790	9 467	10 821	13 339	16 222	18 151	21 064	23 685	26 873	29 141	36 123
15	5.229	5 985	7.261	8.547	10.307	11 721	14 339	17.322	19 311	22 307	24 996	28 259	30 578	37 697
16	5 812	6 614	7 962	9.312	11 152	12 624	15.338	18.418	20 465	23 542	26 296	29 633	32 000	39 252
17	6 408	7.255	8.672	10 085	12.002	13.531	16.338	19 511	21 615	24 769	27 587	30 995	33 409	40 790
18	7 015	7 906	9.390	10 865	12 857	14 440	17 338	20 001	22 760	25 989	28 869	32 346	34 805	42 312
19	7 633	8.567	10 117	11 651	13.716	15.352	18 338	21 689	23 900	27 204	30 144	33 687	36 191	43 820
20	8 260	9 237	10.851	12.443	14 578	16 266	19.337	22 775	25 038	28 412	31 410	35 020	37 566	45.315
21	8.897	9 915	11 591	13.240	15.445	17 182	20 337	23 858	26 171	29 615	32 671	36 343	38 932	46.797
22	9 542	10 600	12.338	14 041	16.314						33 924	37 659	40 289	48 268
23	10 196	11.293	13 091	14.848	17 187						35 172	38 968	41 638	49 728
24	10 856	11 992	13 848	15 659	18.062	19 943	23.337							
25	11.524	12.697	14.611	16 473	18 940	20.867	24 337							
26	12 196	13.409	15.379	17.292	19.820	21 792	25.836	29.246	31 795	35.563	38 885	42 856	45 642	54 052
27	12 879	14.125	16 151	18 114	20 703									
28	13 565	14.847	16 928	18 939	21 588									
29	14.256	15.574	17 708	19 768	22 475									
30	14.953	16.306	18.493	20.599	23 364	25 508	29 336	33 530	36 250	40 256	43 773	47 962	50 892	59 703

For large values of n compute $\sqrt{2\chi^2}$, the distribution of which is approximately normal around a mean of $\sqrt{2n-1}$ with $\sigma = 1$. P is the ratio of one tail of the normal distribution to the area under the entire curve.

This table is taken by consent from *Statistical Tables for Biological, Agricultural, and Medical Research*, by Prof. R. A. Fisher and F. Yates, published @ 12/6 by Oliver and Boyd, Edinburgh.

A detailed table of the probability of various values of χ^2 for one degree of freedom is given in G. U. Yule and M. G. Kendall, *An Introduction to the Theory of Statistics*, 11th edition, pp. 534–535, Charles Griffin and Co., London, 1937.
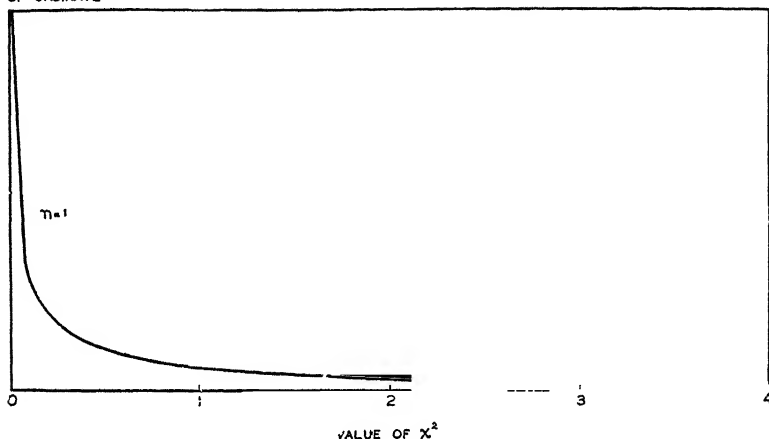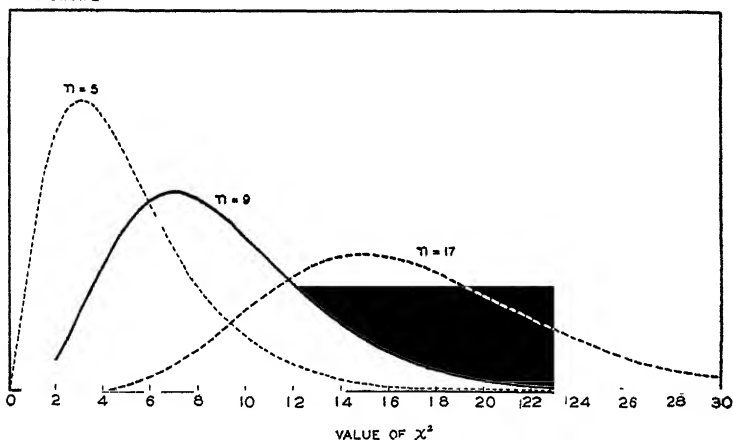
Distribution of χ^2 for $n = 1$, $n = 5$, $n = 9$, and $n = 17$. The maximum ordinate is at $\chi^2 = n - 2$ except when $n = 1$. When $n = 1$, the maximum ordinate is at $\chi^2 = 0$. When $n = 1$, there is 4.55 per cent of the curve beyond $\chi^2 = 4$. Beyond $\chi^2 = 30$ there is .0015 of one per cent of the curve when $n = 5$; 0439 of one per cent of the curve when $n = 9$; 2.6345 per cent of the curve when $n = 17$. The two charts have been drawn to different scales. If the vertical axis of the upper chart is expanded to approximately 20 times its length and the horizontal axis is contracted to about one-eighth of its length, the curves will be roughly comparable as to area.

883

APPENDIX J

Values of $F_2\left(\dfrac{x}{\sigma}\right)$

For Use in Fitting Curves of the Type

$$Y_c = \frac{Ni}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}} - \left\{ \frac{Ni}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}} \left[\frac{\alpha_3}{2}\left(\frac{x}{\sigma} - \frac{x^3}{3\sigma^3}\right) \right] \right\} = \frac{Ni}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}} \left[1 - \frac{\alpha_3}{2}\left(\frac{x}{\sigma} - \frac{x^3}{3\sigma^J}\right) \right]$$

$\frac{x}{\sigma}$	.00	01	.02	03	04	05	.06	07	08	.09
.0	00000	00001	00004	00009	00016	.00025	00036	00049	.00064	00081
.1	00099	00120	.00143	00167	.00194	00222	.00253	00285	00319	.00355
.2	.00392	00432	00473	.00516	00561	00607	00656	00705	.00757	00810
.3	00865	00921	00979	01038	01099	.01161	01225	.01290	01356	01424
.4	01493	01564	01635	01708	01782	01857	01933	02011	.02089	02168
.5	02248	02329	02411	.02494	02578	02662	02748	.02833	.02920	.03007
.6	03095	03183	03272	03361	03450	.03540	03631	.03721	03812	.03904
.7	03995	.04086	04178	04270	04362	04453	.04545	.04637	.04728	04820
.8	04911	05002	05093	05183	05274	05363	05453	05542	05631	05719
9	05806	.05894	.05980	.06066	06152	06236	06320	06404	.06486	.06568
1 0	06649	06729	06809	06887	06965	07042	07118	07193	07267	.07340
1 1	07412	.07483	07552	07621	07689	07756	07822	07886	07950	.08012
1 2	08073	08133	08192	08250	08306	08361	08416	08468	08520	08571
1.3	08620	08668	08715	08760	08805	08848	08890	08930	08970	09008
1 4	09045	09080	09115	09148	09180	09211	09241	.09269	.09296	.09322
1 5	09347	09371	09394	09415	.09435	09454	09472	09489	09505	09519
1 6	09533	09546	09557	09567	09577	09585	09592	09599	.09604	.09608
1 7	09612	09614	09616	09616	09616	09615	09613	09610	09606	09602
1 8	.09597	09590	09584	09576	09568	09559	09549	.09539	09527	09516
1 9	09503	09490	.09477	.09463	09448	09433	09417	09401	09384	09366
2 0	.09349	.09330	09312	.09293	09273	09253	09233	.09213	09192	.09170
2 1	.09149	09127	09105	09082	09060	09037	09014	.08991	08967	.08943
2 2	.08919	08895	08871	.08847	.08823	08798	08774	08749	08724	.08699
2 3	08674	08650	08625	08600	.08575	08550	08525	08500	.08475	08450
2 4	.08426	08401	.08376	08352	08327	.08303	.08279	08255	08231	08207
2 5	.08183	08159	.08136	08112	08089	08066	.08043	08020	.07998	.07975
2 6	.07953	07931	07909	07888	.07866	.07845	07824	07803	.07782	.07762
2 7	07742	.07722	07702	07682	07663	.07644	.07625	07606	07588	07569
2 8	.07551	.07534	07516	07499	07482	.07465	.07448	.07432	.07416	.07400
2 9	07384	.07369	.07354	07339	07324	.07309	.07295	.07281	.07267	.07254
3.0	07240									
3 1	07118									
3 2	07016									
3 3	06933									
3 4	06866									
3 5	06813									
3 6	06771									
3 7	.06739									
3 8	06714									
3 9	06696									
4.0	06683									

From W. A. Shewhart, *Economic Control of Quality of Manufactured Product*, p. 91, D. Van Nostrand Company, Inc., New York, 1931. Courtesy of D. Van Nostrand Company, Inc. and The Bell Telephone Laboratories.

For values of $F_2\left(\dfrac{x}{\sigma}\right)$ beyond the range shown above, use the expression $F_2\left(\dfrac{x}{\sigma}\right)$

$$= \frac{1}{6\sqrt{2\pi}} \left\{ 1 - \left[1 - \left(\frac{x}{\sigma}\right)^2 \right] e^{\frac{-x^2}{2\sigma^2}} \right\} = \frac{1}{15.036} \left\{ 1 - \left[1 - \left(\frac{x}{\sigma}\right)^2 \right] e^{\frac{-x^2}{2\sigma^2}} \right\}.$$ The values of $e^{\frac{-x^2}{2\sigma^2}}$

may be conveniently read from the table of ordinates of the normal curve, Appendix D, or from a more extensive table in Karl Pearson, *Tables for Statisticians and Biometricians*, pp. 2–8, The University Press, Cambridge, England, 1914. The values for z shown in the latter table yield $e^{\frac{-x^2}{2\sigma^2}}$ when multiplied by 2.5066.

APPENDIX K

Flexible Calendar of Working Days

Calendar Days, Sundays, Saturdays, and Holidays, by Months, 1898-1950

The first row for each year gives the number of Sundays (in parentheses) and Saturdays [in brackets] in each month. The second row shows the occurrence of holidays. Holidays occurring on Sundays are enclosed in parentheses; those on Saturdays are enclosed in brackets. For convenience of reference the years are arranged in columns according to decades with some overlapping between columns. In all there are 14 distinct calendar patterns referred to in this table by code number. For information concerning the states in which specific holidays are observed, see *The World Almanac* (published annually by the New York *World Telegram*, New York City).

Following is a key to the symbols used on the calendar:

N New Year's Day—January 1.
L Lincoln's Birthday—February 12.
W Washington's Birthday—February 22.
F Good Friday.
E Easter.
M Memorial Day—May 30.
J Independence Day—July 4.

D Labor Day—First Monday in September.
C Columbus Day—October 12.
V Election Day—First Tuesday after First Monday in November.
A Armistice Day—November 11 (beginning 1918).
T Thanksgiving Day.
X Christmas Day—December 25.

Perpetual calendar / day-of-week reference table (rotated). The column headings (left to right) are: **Year** (five sub-columns), **Code Number**, then the months **Jan 31, Feb 28, Mar 31, Apr 30, May 31, Jun 30, Jul 31, Aug 31, Sep 30, Oct 31, Nov 30, Dec 31**. Each month cell shows a pair of figures such as "(4) [5]" with an identifying letter beneath (N for January; W, L for February; F, E for April; M for May; J for July; D for September; C for October; V, A, T for November; X for December).

Year					Code Number	Jan 31 (N)	Feb 28 (L, W)	Mar 31	Apr 30 (F, E)	May 31 (M)	Jun 30	Jul 31 (J)	Aug 31	Sep 30 (D)	Oct 31 (C)	Nov 30 (V A T)	Dec 31 (X)
1898	1910fe	1921fe	1927	1938	I	(4)[5]	(4)[4]	(4)	(4)[5]	(5)[4]	(4)	(5)[5]	(4)	(4)[4]	(4)[4]	(4)[4]	(4)[5]
.	1911	.	1928*	..	II	(5)[4]	(4)[4]	(4)	(5)[4]	(4)[4]	(4)[5]	(5)[4]	(4)	(5)[5]	(4)[4]	(4)[4]	(5)[5]
1899f	1912*	1922		1939	III	(5)[4]	(4)[4]	(4)[5]	(5)[5]	(4)[4]	(4)	(5)	(4)	(4)[5]	(5)[5]	(4)[4]	(5)
1900†		1923f			IV	(4)[4]	(4)[4]	(4)	(5)[4]	(4)[4]	(4)	(4)	(5)	(4)[5]	(4)[4]	(4)[4]	(5)[4]
.	1912*	.	1929fe	1940*fe	V	(4)[4]	(4)[4]	(5)[5]	(4)[4]	(4)[4]	(5)	(4)[5]	(4)	(5)[4]	(4)[4]	(4)[5]	(5)[4]
1901	1913fe	.	1930	1941	VI	(4)[4]	(4)[4]	(5)	(4)[4]	(4)[5]	(5)	(4)[4]	(5)	(5)[4]	(4)[4]	(4)[4]	(5)
1902fe	.	1924*	.	.	VII	(4)[4]	(4)[4]	(5)	(4)[4]	(4)[5]	(5)	(4)[4]	(5)	(4)[4]	(4)[4]	(7)[7]	(4)
.	1914	1925	1931	1942	VIII	(4)[4]	(4)[4]	(5)	(4)[4]	(4)[4]	(4)	(4)[4]	(4)	(4)[4]	(4)[5]	(4)[4]	(4)
1903	1915	1926	.	1943	IX	(4)[5]	(4)[4]	(5)	(4)[4]	(5)[5]	(4)	(4)[5]	(5)	(4)[4]	(5)[5]	(4)[4]	(4)
.	.	.	1932*fe	.	X	(5)[5]	(4)[4]	(4)	(4)[5]	(5)[5]	(4)	(5)[5]	(5)	(4)[4]	(5)[5]	(4)[5]	(4)[5]
1904*	1916*	.	.	1944**	XI	(5)[5]	(4)[5]	(5)	(4)[4]	(4)[4]	(4)	(5)[5]	(4)	(4)[4]	(5)[4]	(4)[4]	(4)
.	.	1933	.	.	XII	(5)[5]	(4)[4]	(4)	(5)[5]	(4)[4]	(4)	(5)[5]	(4)	(4)[4]	(5)[5]	(4)[4]	(5)[5]
1905	1917	1934f	1945f	.	III	(4)[4]	(4)[4]	(5)	(5)[5]	(4)[4]	(4)	(4)[5]	(5)	(4)[4]	(4)[4]	(4)[4]	(5)
1906	1918fe	1935	1946	.	IV	(4)[4]	(4)[4]	(4)	(5)[4]	(4)[5]	(5)	(3)[4]	(5)	(5)[4]	(4)[4]	(4)[4]	(4)
1907fe	.	1936*	.	.	VI	(4)[4]	(4)[4]	(4)	(4)[4]	(4)[4]	(4)	(4)	(5)	(5)[4]	(4)[5]	(4)[4]	(4)[4]
1908*	1919	.	1947	.	XIII	(4)[4]	(4)[5]	(5)	(4)[4]	(5)[5]	(5)	(4)[4]	(4)	(4)[4]	(4)[4]	(5)[5]	(4)[4]
.	.	1937fe	1948*fe	.	VII	(4)[4]	(4)[4]	(5)	(4)[4]	(4)[5]	(4)	(4)[4]	(5)	(4)[4]	(4)[4]	(5)[5]	(4)[4]
1909	1920*	.	.	.	X	(5)[5]	(4)[4]	(5)	(4)[5]	(5)[5]	(4)	(4)	(5)	(4)[4]	(4)[5]	(4)[4]	[X][4]
.	1921fe	1937fe	1949	.	XIV	(4)[4]	(4)[4]	(4)	(5)[5]	(5)[5]	(4)	(4)[5]	(5)	(4)[4]	(5)	(4)[4]	(4)[4]
1910fe	1922	1938	1938	.	I	(4)[4]	(4)[4]	(4)	(4)[4]	(4)[4]	(4)	(5)[5]	(4)	(4)[4]	(5)[5]	(5)[5]	(5)[5]
1911		1939	1950	.	III	(5)[4]	(4)[4]	(4)	(5)[5]	(4)[4]	(4)	(5)[5]	(4)	(4)[5]	(5)[4]	(4)[4]	(5)

* Leap Year, February has 29 days   † 1900 was not a Leap Year   f Good Friday occurred in March   e Easter occurred in March

APPENDIX L

Brief Table of Sines and Cosines

To find the sine or cosine of any angle greater than 90°, subtract the angle from some integral multiple of 180° and use the table below. Signs to be prefixed to table values are as follows:

Angle (degrees)	Sin	Cos
0–90	+	+
90–180	+	−
180–270	−	−
270–360	−	+

When using this table to determine correlation by Sheppard's method of Unlike Signs (Cos U 1.8°), values of the coefficient are negative when U 1.8° exceeds 90°.

Degree	Sine	Cosine	Degree	Sine	Cosine	Degree	Sine	Cosine
0	.0000	1.0000	30	.5000	.8660	60	.8660	.5000
1	.0175	.9998	31	.5150	.8572	61	.8746	.4848
2	.0349	.9994	32	.5299	.8480	62	.8829	.4695
3	.0523	.9986	33	.5446	.8387	63	.8910	.4540
4	.0698	.9976	34	.5592	.8290	64	.8988	.4384
5	.0872	.9962	35	.5736	.8192	65	.9063	.4226
6	.1045	.9945	36	.5878	.8090	66	.9135	.4067
7	.1219	.9925	37	6018	.7986	67	.9205	.3907
8	.1392	.9903	38	.6157	.7880	68	.9272	.3746
9	.1564	.9877	39	.6293	.7771	69	.9336	.3584
10	.1736	.9848	40	.6428	.7660	70	.9397	.3420
11	.1908	.9816	41	.6561	.7547	71	.9455	.3256
12	.2079	.9781	42	.6691	.7431	72	.9511	3090
13	.2250	.9744	43	.6820	.7314	73	.9563	2924
14	.2419	.9703	44	.6947	.7193	74	.9613	.2756
15	.2588	.9659	45	.7071	.7071	75	.9659	.2588
16	.2756	.9613	46	.7193	.6947	76	.9703	.2419
17	.2924	.9563	47	.7314	.6820	77	.9744	.2250
18	.3090	.9511	48	.7431	.6691	78	.9781	.2079
19	.3256	.9455	49	.7547	.6561	79	.9816	.1908
20	.3420	.9397	50	.7660	.6428	80	.9848	.1736
21	.3584	.9336	51	.7771	.6293	81	.9877	.1564
22	.3746	.9272	52	.7880	.6157	82	.9903	.1392
23	.3907	.9205	53	.7986	.6018	83	.9925	.1219
24	.4067	.9135	54	.8090	.5878	84	.9945	.1045
25	.4226	.9063	55	.8192	.5736	85	.9962	.0872
26	.4384	.8988	56	.8290	.5592	86	.9976	.0698
27	.4540	.8910	57	.8387	.5446	87	.9986	.0523
28	.4695	.8829	58	.8480	.5299	88	.9994	.0349
29	.4848	.8746	59	.8572	.5150	89	.9998	.0175
						90	1.0000	.0000

APPENDIX M

Sums of First Six Powers of First 50 Natural Numbers

A table of the sums of the squares of the first 100 natural numbers may be found in Croxton and Cowden, *Practical Business Statistics*, page 494   A table of the sums of the first 7 powers of the first 100 natural numbers may be found in Pearson, *Tables for Statisticians and Biometricians*, pages 40–41.   The sums of the powers of the first M natural numbers may also be computed by the following formulae:

$$\sum_1^M X = \frac{M(M+1)}{2}$$

$$\sum_1^M X^2 = \left(\frac{2M+1}{3}\right)\sum_1^M X$$

$$\sum_1^M X^3 = \left(\sum_1^M X\right)^2$$

$$\sum_1^M X^4 = \left(\frac{3M^2+3M-1}{5}\right)\sum_1^M X^2$$

$$\sum_1^M X^5 = \left(\frac{2M^2+2M-1}{3}\right)\sum_1^M X^3$$

$$\sum_1^M X^6 = \left(\frac{3M^4+6M^3-3M+1}{7}\right)\sum_1^M X^2$$

M is the highest value of X used in the computation table   When the X origin is taken at the center of the X values, it is necessary to multiply the summation value of this table by 2   In this case N as used in the normal equations is $2M+1$

M	$\sum_1^M X$	$\sum_1^M X^2$	$\sum_1^M X^3$	$\sum_1^M X^4$	$\sum_1^M X^5$	$\sum_1^M X^6$
1	1	1	1	1	1	1
2	3	5	9	17	33	65
3	6	14	36	98	276	794
4	10	30	100	354	1 300	4 890
5	15	55	225	979	4 425	20 515
6	21	91	441	2 275	12 201	67 171
7	28	140	784	4 676	29 008	184 820
8	36	204	1 296	8 772	61 776	446 964
9	45	285	2 025	15 333	120 825	978 405
10	55	385	3 025	25 333	220 825	1 978 405
11	66	506	4 356	39 974	381 874	3 749 966
12	78	650	6 084	60 710	630 708	6 735 950
13	91	819	8 281	89 271	1 002 001	11 562 759
14	105	1 015	11 025	127 687	1 539 825	19 092 295
15	120	1 240	14 400	178 312	2 299 200	30 482 920
16	136	1 496	18 496	234 848	3 347 776	47 260 136
17	153	1 785	23 409	327 369	4 767 633	71 397 705
18	171	2 109	29 241	432 345	6 657 201	105 409 929
19	190	2 470	36 100	562 666	9 133 300	152 455 810
20	210	2 870	44 100	722 666	12 333 300	216 455 810
21	231	3 311	53 361	917 147	16 417 401	302 221 931
22	253	3 795	64 009	1 151 403	21 571 033	415 601 835
23	276	4 324	76 176	1 431 244	28 007 376	563 637 724
24	300	4 900	90 000	1 763 020	35 970 000	754 740 700
25	325	5 525	105 625	2 153 645	45 735 625	998 881 325
26	351	6 201	123 201	2 610 621	57 617 001	1 307 797 101
27	378	6 930	142 884	3 142 062	71 965 908	1 695 217 590
28	406	7 714	164 836	3 756 718	89 176 276	2 177 107 894
29	435	8 555	189 225	4 463 999	109 687 425	2 771 931 215
30	465	9 455	216 225	5 273 999	133 987 425	3 500 931 215
31	496	10 416	246 016	6 197 520	162 616 576	4 388 434 896
32	528	11 440	278 784	7 246 096	196 171 008	5 462 176 720
33	561	12 529	314 721	8 432 017	235 306 401	6 753 644 689
34	595	13 685	354 025	9 768 353	280 741 825	8 298 449 105
35	630	14 910	396 900	11 268 978	333 263 700	10 136 714 730
36	666	16 206	443 556	12 948 594	393 729 876	12 313 497 066
37	703	17 575	494 209	14 822 755	463 073 833	14 879 223 475
38	741	19 019	549 081	16 907 891	542 309 001	17 890 159 859
39	780	20 540	608 400	19 221 332	632 533 200	21 408 903 620
40	820	22 140	672 400	21 781 332	734 933 200	25 504 903 620
41	861	23 821	741 321	24 607 093	850 789 401	30 255 007 861
42	903	25 585	815 409	27 718 789	981 480 633	35 744 039 605
43	946	27 434	894 916	31 137 590	1 128 489 076	42 065 402 654
44	990	29 370	980 100	34 885 686	1 293 405 300	49 321 716 510
45	1 035	31 395	1 071 225	38 986 311	1 477 933 425	57 625 482 135
46	1 081	33 511	1 168 561	43 463 767	1 683 896 401	67 099 779 021
47	1 128	35 720	1 272 384	48 343 448	1 913 241 408	77 878 994 360
48	1 176	38 024	1 382 976	53 651 864	2 168 045 376	90 109 584 824
49	1 225	40 425	1 500 625	59 416 665	2 450 520 625	103 950 872 025
50	1 275	42 925	1 625 625	65 666 665	2 763 020 625	119 575 872 025

APPENDIX N

Sums of the First Six Powers of the First 50 Odd Natural Numbers

A table of the sums of the squares of the first 100 odd natural numbers may be found in Croxton and Cowden, *Practical Business Statistics*, page 495. A table of the sums of the first 6 powers of the first 100 odd natural numbers is given in Ross "Formulae for Facilitating Computations in Time Series Analysis," *Journal of the American Statistical Association*, March 1925, pages 75–79.

The sums of the powers of the first M_o odd natural numbers may be computed by the following formulae:

$$\sum_1^{M_o} X_o = M_o^2$$

$$\sum_1^{M_o} X_o^4 = \left(\frac{12M_o^2 - 7}{5}\right)\sum_1^{M_o} X_o^2$$

$$\sum_1^{M_o} X_o^2 = \frac{4M_o^3 - M_o}{3}$$

$$\sum_1^{M_o} X_o^5 = \left(\frac{16M_o^4 - 20M_o^2 + 7}{3}\right)\sum_1^{M_o} X_o$$

$$\sum_1^{M_o} X_o^3 = (2M_o^2 - 1)\sum_1^{M_o} X_o$$

$$\sum_1^{M_o} X_o^6 = \left(\frac{48M_o^4 - 72M_o^2 + 31}{7}\right)\sum_1^{M_o} X_o^2$$

M is the highest value of X_o (odd value of X) used in the computation table. M_o may be ascertained by reference to the first two columns of this appendix or from the expression $M_o = \frac{M+1}{2}$. When the X origin is taken at the center of the X_o values, it is necessary to multiply the summation value of this table by 2. In this case N as used in the normal equation is $2M_o$.

M	M_o	$\sum_{1}^{M_o} X_0$	$\sum_{1}^{M_o} X_0^2$	$\sum_{1}^{M_o} X_0^3$	$\sum_{1}^{M_o} X_0^4$	$\sum_{1}^{M_o} X_0^5$	$\sum_{1}^{M_o} X_0^6$
1	1	1	1	1	1	1	1
3	2	4	10	28	82	244	730
5	3	9	35	153	707	3 369	16 355
7	4	16	84	496	3 108	20 176	134 004
9	5	25	165	1 225	9 669	79 225	665 445
11	6	36	286	2 556	24 310	240 276	2 437 006
13	7	49	455	4 753	52 871	611 569	7 263 815
15	8	64	680	8 128	103 496	1 370 944	18 654 440
17	9	81	969	13 041	187 017	2 790 801	42 792 009
19	10	100	1 330	19 900	317 338	5 266 900	89 837 890
21	11	121	1 771	29 161	511 819	9 351 001	175 604 011
23	12	144	2 300	41 328	791 660	15 787 344	323 639 900
25	13	169	2 925	56 953	1 182 285	25 552 969	567 780 525
27	14	196	3 654	76 636	1 713 726	39 901 876	955 201 014
29	15	225	4 495	101 025	2 421 007	60 413 025	1 550 024 335
31	16	256	5 456	130 816	3 344 528	89 042 176	2 437 528 016
33	17	289	6 545	166 753	4 530 449	128 177 569	3 728 995 985
35	18	324	7 770	209 628	6 031 074	180 699 444	5 567 261 610
37	19	361	9 139	260 281	7 905 235	250 043 401	8 132 988 019
39	20	400	10 660	319 600	10 218 676	340 267 600	11 651 731 780
41	21	441	12 341	388 521	13 044 437	456 123 801	16 401 836 021
43	22	484	14 190	468 028	16 463 238	603 132 244	22 723 199 070
45	23	529	16 215	559 153	20 563 863	787 660 369	31 026 964 695
47	24	576	18 424	662 976	25 443 544	1 017 005 376	41 806 180 024
49	25	625	20 825	780 625	31 208 345	1 299 480 625	55 647 467 225
51	26	676	23 426	913 276	37 973 546	1 644 505 876	73 243 755 026
53	27	729	26 235	1 062 153	45 864 027	2 062 701 369	95 408 116 155
55	28	784	29 260	1 228 528	55 014 652	2 565 985 744	123 088 756 780
57	29	841	32 509	1 413 721	65 570 653	3 167 677 801	157 385 204 029
59	30	900	35 990	1 619 100	77 688 014	3 882 602 100	199 565 737 67C
61	31	961	39 711	1 846 081	91 533 855	4 727 198 401	251 086 112 031
63	32	1 024	43 680	2 096 128	107 286 816	5 719 634 944	313 609 614 240
65	33	1 089	47 905	2 370 753	125 137 441	6 879 925 569	389 028 504 865
67	34	1 156	52 394	2 671 516	145 288 562	8 230 050 676	479 486 887 034
69	35	1 225	57 155	3 000 025	167 955 683	9 794 082 025	537 405 050 115
71	36	1 296	62 196	3 357 936	193 367 364	11 598 311 376	715 505 334 036
73	37	1 369	67 525	3 746 953	221 765 605	13 671 382 969	866 839 560 325
75	38	1 441	73 150	4 168 828	253 406 230	16 044 429 844	1 044 818 075 950
77	39	1 521	79 079	4 625 361	288 559 271	18 751 214 001	1 253 240 456 039
79	40	1 600	85 320	5 118 400	327 509 352	21 828 270 400	1 496 327 911 560
81	41	1 681	91 881	5 649 841	370 556 073	25 315 054 801	1 778 757 448 041
83	42	1 764	98 770	6 221 628	418 014 394	29 254 095 444	2 105 697 821 410
85	43	1 849	105 995	6 835 753	470 215 019	33 691 148 569	2 482 847 337 035
87	44	1 936	113 564	7 494 256	527 504 780	38 675 357 776	2 916 473 538 044
89	45	2 025	121 485	8 199 225	590 247 021	44 259 417 225	3 413 454 829 005
91	46	2 116	129 766	8 952 796	658 821 982	50 499 738 676	3 981 324 081 046
93	47	2 209	138 415	9 757 153	733 627 183	57 456 622 369	4 628 314 264 495
95	48	2 304	147 440	10 614 528	815 077 808	65 194 431 744	5 363 406 155 120
97	49	2 401	156 849	11 527 201	903 607 089	73 781 772 001	6 196 378 160 049
99	50	2 500	166 650	12 497 500	999 666 690	83 291 672 500	7 137 858 309 450

APPENDIX O

Squares, Square Roots, and Reciprocals
1-1000

No	Square	Square Root	Reciprocal	No.	Square	Square Root	Reciprocal
1	1	1.0000000	1.000000000	51	26 01	7.1414284	.019607843
2	4	1.4142136	0.500000000	52	27 04	7.2111026	019230769
3	9	1.7320508	.333333333	53	28 09	7.2801099	.018867925
4	16	2.0000000	.250000000	54	29 16	7.3484692	.018518519
5	25	2.2360680	.200000000	55	30 25	7.4161985	.018181818
6	36	2.4494897	.166666667	56	31 36	7.4833148	.017857143
7	49	2.6457513	.142857143	57	32 49	7.5498344	.017543860
8	64	2 8284271	.125000000	58	33 64	7.6157731	.017241379
9	81	3.0000000	.111111111	59	34 81	7.6811457	.016949153
10	1 00	3.1622777	.100000000	60	36 00	7.7459667	.016666667
11	1 21	3 3166248	.090909091	61	37 21	7.8102497	.016393443
12	1 44	3.4641016	.083333333	62	38 44	7.8740079	.016129032
13	1 69	3.6055513	.076923077	63	39 69	7.9372539	.015873016
14	1 96	3.7416574	.071428571	64	40 96	8 0000000	.015625000
15	2 25	3.8729833	.066666667	65	42 25	8.0622577	.015384615
16	2 56	4.0000000	.062500000	66	43 56	8 1240384	.015151515
17	2 89	4.1231056	.058823529	67	44 89	8.1853528	.014925373
18	3 24	4.2426407	.055555556	68	46 24	8.2462113	.014705882
19	3 61	4.3588989	.052631579	69	47 61	8 3066239	.014492754
20	4 00	4.4721360	.050000000	70	49 00	8.3666003	.014285714
21	4 41	4.5825757	.047619048	71	50 41	8.4261498	.014084507
22	4 84	4 6904158	.045454545	72	51 84	8.4852814	013888889
23	5 29	4.7958315	.043478261	73	53 29	8.5440037	.013698630
24	5 76	4.8989795	.041666667	74	54 76	8.6023253	.013513514
25	6 25	5 0000000	.040000000	75	56 25	8 6602540	.013333333
26	6 76	5 0990195	.038461538	76	57 76	8 7177979	.013157895
27	7 29	5.1961524	.037037037	77	59 29	8.7749644	.012987013
28	7 84	5.2915026	.035714286	78	60 84	8 8317609	.012820513
29	8 41	5 3851648	.034482759	79	62 41	8.8881944	.012658228
30	9 00	5.4772256	.033333333	80	64 00	8.9442719	.012500000
31	9 61	5.5677644	.032258065	81	65 61	9.0000000	.012345679
32	10 24	5 6568542	.031250000	82	67 24	9.0553851	.012195122
33	10 89	5.7445626	.030303030	83	68 89	9.1104336	.012048193
34	11 56	5.8309519	.029411765	84	70 56	9.1651514	.011904762
35	12 25	5.9160798	.028571429	85	72 25	9.2195445	.011764706
36	12 96	6.0000000	.027777778	86	73 96	9.2736185	.011627907
37	13 69	6.0827625	.027027027	87	75 69	9.3273791	.011494253
38	14 44	6.1644140	.026315789	88	77 44	9.3808315	.011363636
39	15 21	6.2449980	.025641026	89	79 21	9.4339811	.011235955
40	16 00	6.3245553	.025000000	90	81 00	9.4868330	.011111111
41	16 81	6.4031242	.024390244	91	82 81	9.5393920	.010989011
42	17 64	6.4807407	.023809524	92	84 64	9.5916630	.010869565
43	18 49	6.5574385	.023255814	93	86 49	9.6436508	010752688
44	19 36	6.6332496	.022727273	94	88 36	9.6953597	.010638298
45	20 25	6.7082039	.022222222	95	90 25	9.7467943	010526316
46	21 16	6.7823300	.021739130	96	92 16	9.7979590	.010416667
47	22 09	6.8556546	.021276596	97	94 09	9.8488578	.010309278
48	23 04	6.9282032	.020833333	98	96 04	9 8994949	.010204082
49	24 01	7.0000000	.020408163	99	98 01	9.9498744	.010101010
50	25 00	7.0710678	.020000000	100	1 00 00	10.0000000	.010000000

No	Square	Square Root	Reciprocal .00	No.	Square	Square Root	Reciprocal 00
101	1 02 01	10 0498756	9900990	151	2 28 01	12 2882057	6622517
102	1 04 04	10 0995049	9803922	152	2 31 04	12 3288280	6578947
103	1 06 09	10.1488916	9708738	153	2 34 09	12 3693169	6535948
104	1 08 16	10.1980390	9615385	154	2 37 16	12 4096736	6493506
105	1 10 25	10.2469508	9523810	155	2 40 25	12 4498996	6451613
106	1 12 36	10.2956301	9433962	156	2 43 36	12 4899960	6410256
107	1 14 49	10.3440804	9345794	157	2 46 49	12 5299641	6369427
108	1 16 64	10.3923048	9259259	158	2 49 64	12 5698051	6329114
109	1 18 81	10.4403065	9174312	159	2 52 81	12 6095202	6289308
110	1 21 00	10.4880885	9090909	160	2 56 00	12.6491106	6250000
111	1 23 21	10.5356538	9009009	161	2 59 21	12.6885775	6211180
112	1 25 44	10.5830052	8928571	162	2 62 44	12.7279221	6172840
113	1 27 69	10.6301458	8849558	163	2 65 69	12 7671453	6131969
114	1 29 96	10 6770783	8771930	164	2 68 96	12 8062485	6097561
115	1 32 25	10.7238053	8695652	165	2 72 25	12.8452326	6060606
116	1 34 56	10.7703296	8620690	166	2 75 56	12 8840987	6024096
117	1 36 89	10.8166538	8547009	167	2 78 89	12 9228480	5988024
118	1 39 24	10.8627805	8474576	168	2 82 24	12.9614814	5952381
119	1 41 61	10.9087121	8403361	169	2 85 61	13 0000000	5917160
120	1 44 00	10.9544512	8333333	170	2 89 00	13 0384048	5882353
121	1 46 41	11.0000000	8264463	171	2 92 41	13 0766968	5847953
122	1 48 84	11.0453610	8196721	172	2 95 84	13 1148770	5813953
123	1 51 29	11.0905365	8130081	173	2 99 29	13 1529464	5780347
124	1 53 76	11.1355287	8064516	174	3 02 76	13.1909060	5747126
125	1 56 25	11.1803399	8000000	175	3 06 25	13 2287566	5714286
126	1 58 76	11.2249722	7936508	176	3 09 76	13 2664992	5681818
127	1 61 29	11.2694277	7874016	177	3 13 29	13.3041347	5649718
128	1 63 84	11.3137085	7812500	178	3 16 84	13.3416641	5617978
129	1 66 41	11.3578167	7751938	179	3 20 41	13.3790882	5586592
130	1 69 00	11.4017543	7692308	180	3 24 00	13.4164079	5555556
131	1 71 61	11.4455231	7633588	181	3 27 61	13.4536240	5524862
132	1 74 24	11.4891253	7575758	182	3 31 24	13.4907376	5494505
133	1 76 89	11.5325626	7518797	183	3 34 89	13.5277493	5464481
134	1 79 56	11.5758369	7462687	184	3 38 56	13.5646600	54347S3
135	1 82 25	11.6189500	7407407	185	3 42 25	13 6014705	5405405
136	1 84 96	11.6619038	7352941	186	3 45 96	13.6381817	5376344
137	1 87 69	11.7046999	7299270	187	3 49 69	13 6747943	5347594
138	1 90 44	11.7473401	7246377	188	3 53 44	13.7113092	5319149
139	1 93 21	11.7898261	7194245	189	3 57 21	13.7477271	5291005
140	1 96 00	11.8321596	7142857	190	3 61 00	13.7840488	5263158
141	1 98 81	11.8743422	7092199	191	3 64 81	13 8202750	5235602
142	2 01 64	11.9163753	7042254	192	3 68 64	13.8564065	5208333
143	2 04 49	11.9582607	6993007	193	3 72 49	13.8924440	5181347
144	2 07 36	12 0000000	6944444	194	3 76 36	13.9283883	5154639
145	2 10 25	12.0415946	6896552	195	3 80 25	13.9642400	5128205
146	2 13 16	12 0830460	6849315	196	3 84 16	14.0000000	5102041
147	2 16 09	12.1243557	6802721	197	3 88 09	14 0356688	5076142
148	2 19 04	12.1655251	6756757	198	3 92 04	14 0712473	5050505
149	2 22 01	12 2065556	6711409	199	3 96 01	14 1067360	5025126
150	2 25 00	12.2474487	6666667	200	4 00 00	14 1421356	5000000

No.	Square	Square Root	Reciprocal .00	No.	Square	Square Root	Reciprocal 00
201	4 04 01	14.1774469	4975124	251	6 30 01	15.8429795	3984064
202	4 08 04	14.2126704	4950495	252	6 35 04	15.8745079	3968254
203	4 12 09	14.2478068	4926108	253	6 40 09	15.9059737	3952569
204	4 16 16	14.2828569	4901961	254	6 45 16	15.9373775	3937008
205	4 20 25	14.3178211	4878049	255	6 50 25	15.9687194	3921569
206	4 24 36	14.3527001	4854369	256	6 55 36	16.0000000	3906250
207	4 28 49	14.3874946	4830918	257	6 60 49	16.0312195	3891051
208	4 32 64	14.4222051	4807692	258	6 65 64	16.0623784	3875969
209	4 36 81	14.4568323	4784689	259	6 70 81	16.0934769	3861004
210	4 41 00	14.4913767	4761905	260	6 76 00	16.1245155	3846154
211	4 45 21	14.5258390	4739336	261	6 81 21	16.1554944	3831418
212	4 49 44	14.5602198	4716981	262	6 86 44	16.1864141	3816794
213	4 53 69	14.5945195	4694836	263	6 91 69	16.2172747	3802281
214	4 57 96	14.6287388	4672897	264	6 96 96	16.2480768	3787879
215	4 62 25	14.6628783	4651163	265	7 02 25	16.2788206	3773585
216	4 66 56	14.6969385	4629630	266	7 07 56	16.3095064	3759398
217	4 70 89	14.7309199	4608295	267	7 12 89	16 3401346	3745318
218	4 75 24	14.7648231	4587156	268	7 18 24	16.3707055	3731343
219	4 79 61	14.7986486	4566210	269	7 23 61	16.4012195	3717472
220	4 84 00	14.8323970	4545455	270	7 29 00	16.4316767	3703704
221	4 88 41	14.8660687	4524887	271	7 34 41	16.4620776	3690037
222	4 92 84	14.8996644	4504505	272	7 39 84	16.4924225	3676471
223	4 97 29	14.9331845	4484305	273	7 45 29	16.5227116	3663004
224	5 01 76	14.9666295	4464286	274	7 50 76	16.5529454	3649635
225	5 06 25	15 0000000	4444444	275	7 56 25	16 5831240	3636364
226	5 10 76	15.0332964	4424779	276	7 61 76	16 6132477	3623188
227	5 15 29	15.0665192	4405286	277	7 67 29	16.6433170	3610108
228	5 19 84	15.0996689	4385965	278	7 72 84	16.6733320	3597122
229	5 24 41	15.1327460	4366812	279	7 78 41	16.7032931	3584229
230	5 29 00	15.1657509	4347826	280	7 84 00	16.7332005	3571429
231	5 33 61	15.1986842	4329004	281	7 89 61	16.7630546	3558719
232	5 38 24	15.2315462	4310345	282	7 95 24	16.7928556	3546099
233	5 42 89	15.2643375	4291845	283	8 00 89	16.8226038	3533569
234	5 47 56	15.2970585	4273504	284	8 06 56	16 8522995	3521127
235	5 52 25	15.3297097	4255319	285	8 12 25	16.8819430	3508772
236	5 56 96	15.3622915	4237288	286	8 17 96	16.9115345	3496503
237	5 61 69	15.3948043	4219409	287	8 23 69	16.9410743	3484321
238	5 66 44	15.4272486	4201681	288	8 29 44	16.9705627	3472222
239	5 71 21	15.4596248	4184100	289	8 35 21	17.0000000	3460208
240	5 76 00	15.4919334	4166667	290	8 41 00	17.0293864	3448276
241	5 80 81	15.5241747	4149378	291	8 46 81	17.0587221	3436426
242	5 85 64	15.5563492	4132231	292	8 52 64	17.0880075	3424658
243	5 90 49	15.5884573	4115226	293	8 58 49	17.1172428	3412969
244	5 95 36	15.6204994	4098361	294	8 64 36	17.1464282	3401361
245	6 00 25	15.6524758	4081633	295	8 70 25	17.1755640	3389831
246	6 05 16	15.6843871	4065041	296	8 76 16	17.2046505	3378378
247	6 10 09	15.7162336	4048583	297	8 82 09	17.2336879	3367003
248	6 15 04	15.7480157	4032258	298	8 88 04	17.2626765	3355705
249	6 20 01	15.7797338	4016064	299	8 94 01	17.2916165	3344482
250	6 25 00	15 8113883	4000000	300	9 00 00	17.3205081	3333333

No.	Square	Square Root	Reciprocal .00	No.	Square	Square Root	Reciprocal .00
301	9 06 01	17.3493516	3322259	351	12 32 01	18.7349940	2849003
302	9 12 04	17.3781472	3311258	352	12 39 04	18.7616630	2840909
303	9 18 09	17.4068952	3300330	353	12 46 09	18.7882942	2832861
304	9 24 16	17.4355958	3289474	354	12 53 16	18.8148877	2824859
305	9 30 25	17.4642492	3278689	355	12 60 25	18.8414437	2816901
306	9 36 36	17.4928557	3267974	356	12 67 36	18.8679623	2808989
307	9 42 49	17.5214155	3257329	357	12 74 49	18.8944436	2801120
308	9 48 64	17.5499288	3246753	358	12 81 64	18.9208879	2793296
309	9 54 81	17.5783958	3236246	359	12 88 81	18.9472953	2785515
310	9 61 00	17.6068169	3225806	360	12 96 00	18.9736660	2777778
311	9 67 21	17.6351921	3215434	361	13 03 21	19.0000000	2770083
312	9 73 44	17.6635217	3205128	362	13 10 44	19.0262976	2762431
313	9 79 69	17.6918060	3194888	363	13 17 69	19.0525589	2754821
314	9 85 96	17.7200451	3184713	364	13 24 96	19.0787840	2747253
315	9 92 25	17.7482393	3174603	365	13 32 25	19.1049732	2739726
316	9 98 56	17.7763888	3164557	366	13 39 56	19.1311265	2732240
317	10 04 89	17.8044938	3154574	367	13 46 89	19.1572441	2724796
318	10 11 24	17.8325545	3144654	368	13 54 24	19.1833261	2717391
319	10 17 61	17.8605711	3134796	369	13 61 61	19.2093727	2710027
320	10 24 00	17.8885438	3125000	370	13 69 00	19.2353841	2702703
321	10 30 41	17.9164729	3115265	371	13 76 41	19.2613603	2695418
322	10 36 84	17.9443584	3105590	372	13 83 84	19.2873015	2688172
323	10 43 29	17.9722008	3095975	373	13 91 29	19.3132079	2680965
324	10 49 76	18.0000000	3086420	374	13 98 76	19.3390796	2673797
325	10 56 25	18.0277564	3076923	375	14 06 25	19.3649167	2666667
326	10 62 76	18.0554701	3067485	376	14 13 76	19.3907194	2659674
327	10 69 29	18.0831413	3058104	377	14 21 29	19.4164878	2652520
328	10 75 84	18.1107703	3048780	378	14 28 84	19.4422221	2645503
329	10 82 41	18.1383571	3039514	379	14 36 41	19.4679223	2638522
330	10 89 00	18.1659021	3030303	380	14 44 00	19.4935887	2631579
331	10 95 61	18.1934054	3021148	381	14 51 61	19.5192213	2624672
332	11 02 24	18.2208672	3012048	382	14 59 24	19.5448203	2617801
333	11 08 89	18.2482876	3003003	383	14 66 89	19.5703858	2610966
334	11 15 56	18.2756669	2994012	384	14 74 56	19.5959179	2604167
335	11 22 25	18.3030052	2985075	385	14 82 25	19.6214169	2597403
336	11 28 96	18.3303028	2976190	386	14 89 96	19.6468827	2590674
337	11 35 69	18.3575598	2967359	387	14 97 69	19.6723156	2583979
338	11 42 44	18.3847763	2958580	388	15 05 44	19.6977156	2577320
339	11 49 21	18.4119526	2949853	389	15 13 21	19.7230829	2570694
340	11 56 00	18.4390889	2941176	390	15 21 00	19.7484177	2564103
341	11 62 81	18.4661853	2932551	391	15 28 81	19.7737199	2557545
342	11 69 64	18.4932420	2923977	392	15 36 64	19.7989899	2551020
343	11 76 49	18.5202592	2915452	393	15 44 49	19.8242276	2544529
344	11 83 36	18.5472370	2906977	394	15 52 36	19.8494332	2538071
345	11 90 25	18.5741756	2898551	395	15 60 25	19.8746069	2531646
346	11 97 16	18.6010752	2890173	396	15 68 16	19.8997487	2525253
347	12 04 09	18.6279360	2881844	397	15 76 09	19.9248588	2518892
348	12 11 04	18.6547581	2873563	398	15 84 04	19.9499373	2512563
349	12 18 01	18.6815417	2865330	399	15 92 01	19.9749844	2506266
350	12 25 00	18.7082869	2857143	400	16 00 00	20.0000000	2500000

No.	Square	Square Root	Reciprocal .00	No	Square	Square Root	Reciprocal .00
401	16 08 01	20.0249844	2493766	451	20 34 01	21.2367606	2217295
402	16 16 04	20.0499377	2487562	452	20 43 04	21.2602916	2212389
403	16 24 09	20.0748599	2481390	453	20 52 09	21.2837967	2207506
404	16 32 16	20.0997512	2475248	454	20 61 16	21.3072758	2202643
405	16 40 25	20.1246118	2469136	455	20 70 25	21 3307290	2197802
406	16 48 36	20.1494417	2463054	456	20 79 36	21.3541565	2192982
407	16 56 49	20.1742410	2457002	457	20 88 49	21.3775583	2188184
408	16 64 64	20.1990099	2450980	458	20 97 64	21.4009346	2183406
409	16 72 81	20.2237484	2444988	459	21 06 81	21.4242853	2178649
410	16 81 00	20.2484567	2439024	460	21 16 00	21.4476106	2173913
411	16 89 21	20.2731349	2433090	461	21 25 21	21 4709106	2169197
412	16 97 44	20.2977831	2427184	462	21 34 44	21.4941853	2164502
413	17 05 69	20.3224014	2421308	463	21 43 69	21.5174348	2159827
414	17 13 96	20.3469899	2415459	464	21 52 96	21.5406592	2155172
415	17 22 25	20.3715488	2409639	465	21 62 25	21.5638587	2150538
416	17 30 56	20.3960781	2403846	466	21 71 56	21 5870331	2145923
417	17 38 89	20.4205779	2398082	467	21 80 89	21.6101828	2141328
418	17 47 24	20 4450483	2392344	468	21 90 24	21.6333077	2136752
419	17 55 61	20 4694895	2386635	469	21 99 61	21 6564078	2132196
420	17 64 00	20 4939015	2380952	470	22 09 00	21 6794834	2127660
421	17 72 41	20 5182845	2375297	471	22 18 41	21 7025344	2123142
422	17 80 84	20 5426386	2369668	472	22 27 84	21 7255610	2118644
423	17 89 29	20 5669638	2364066	473	22 37 29	21.7485632	2114165
424	17 97 76	20 5912603	2358491	474	22 46 76	21 7715411	2109705
425	18 06 25	20 6155281	2352941	475	22 56 25	21 7944947	2105263
426	18 14 76	20 6397674	2347418	476	22 65 76	21 8174242	2100840
427	18 23 29	20 6639783	2341920	477	22 75 29	21 8403297	2096436
428	18 31 84	20 6881609	2336449	478	22 84 84	21 8632111	2092050
429	18 40 41	20 7123152	2331002	479	22 94 41	21 8860686	2087683
430	18 49 00	20 7364414	2325581	480	23 04 00	21 9089023	2083333
431	18 57 61	20 7605395	2320186	481	23 13 61	21 9317122	2079002
432	18 66 24	20 7846097	2314815	482	23 23 24	21 9544984	2074689
433	18 74 89	20 8086520	2309469	483	23 32 89	21 9772610	2070393
434	18 83 56	20 8326667	2304147	484	23 42 56	22 0000000	2066116
435	18 92 25	20 8566536	2298851	485	23 52 25	22 0227155	2061856
436	19 00 96	20 8806130	2293578	486	23 61 96	22 0454077	2057613
437	19 09 69	20 9045450	2288330	487	23 71 69	22 0680765	2053388
438	19 18 44	20 9284495	2283105	488	23 81 44	22.0907220	2049180
439	19 27 21	20 9523268	2277904	489	23 91 21	22 1133444	2044990
440	19 36 00	20 9761770	2272727	490	24 01 00	22 1359436	2040816
441	19 44 81	21 0000000	2267574	491	24 10 81	22 1585198	2036660
442	19 53 64	21 0237960	2262443	492	24 20 64	22 1810730	2032520
443	19 62 49	21 0475652	2257336	493	24 30 49	22.2036033	2028398
444	19 71 36	21 0713075	2252252	494	24 40 36	22 2261108	2024291
445	19 80 25	21 0950231	2247191	495	24 50 25	22 2485955	2020202
446	19 89 16	21 1187121	2242152	496	24 60 16	22 2710575	2016129
447	19 98 09	21 1423745	2237136	497	24 70 09	22 2934968	2012072
448	20 07 04	21 1660105	2232143	498	24 80 04	22 3159136	2008032
449	20 16 01	21 1896201	2227171	499	24 90 01	22 3383079	2004008
450	20 25 00	21 2132034	2222222	500	25 00 00	22.3606798	2000000

No.	Square	Square Root	Reciprocal .00	No.	Square	Square Root	Reciprocal .00
501	25 10 01	22 3830293	1996008	551	30 36 01	23.4733892	1814882
502	25 20 04	22.4053565	1992032	552	30 47 04	23 4946802	1811594
503	25 30 09	22.4276615	1988072	553	30 58 09	23 5159520	1808318
504	25 40 16	22.4499443	1984127	554	30 69 16	23.5372046	1805054
505	25 50 25	22.4722051	1980198	555	30 80 25	23.5584380	1801802
506	25 60 36	22.4944438	1976285	556	30 91 36	23.5796522	1798561
507	25 70 49	22 5166605.	1972387	557	31 02 49	23.6008474	1795332
508	25 80 64	22.5388553	1968504	558	31 13 64	23 6220236	1792115
509	25 90 81	22.5610283	1964637	559	31 24 81	23 6431808	1788909
510	26 01 00	22 5831796	1960784	560	31 36 00	23.6643191	1785714
511	26 11 21	22.6053091	1956947	561	31 47 21	23.6854386	1782531
512	26 21 44	22.6274170	1953125	562	31 58 44	23.7065392	1779359
513	26 31 69	22.6495033	1949318	563	31 69 69	23 7276210	1776199
514	26 41 96	22.6715681	1945525	564	31 80 96	23.7486842	1773050
515	26 52 25	22.6936114	1941748	565	31 92 25	23.7697286	1769912
516	26 62 56	22.7156334	1937984	566	32 03 56	23.7907545	1766784
517	26 72 89	22.7376340	1934236	567	32 14 89	23 8117618	1763668
518	26 83 24	22.7596134	1930502	568	32 26 24	23 8327506	1760563
519	26 93 61	22.7815715	1926782	569	32 37 61	23.8537209	1757469
520	27 04 00	22 8035085	1923077	570	32 49 00	23 8746728	1754386
521	27 14 41	22 8254244	1919386	571	32 60 41	23.8956063	1751313
522	27 24 84	22.8473193	1915709	572	32 71 84	23 9165215	1748252
523	27 35 29	22.8691933	1912046	573	32 83 29	23.9374184	1745201
524	27 45 76	22.8910463	1908397	574	32 94 76	23.9582971	1742160
525	27 56 25	22.9128785	1904762	575	33 06 25	23 9791576	1739130
526	27 66 76	22 9346899	1901141	576	33 17 76	24 0000000	1736111
527	27 77 29	22.9564806	1897533	577	33 29 29	24.0208243	1733102
528	27 87 84	22.9782506	1893939	578	33 40 84	24 0416306	1730104
529	27 98 41	23.0000000	1890359	579	33 52 41	24.0624188	1727116
530	28 09 00	23.0217289	1886792	580	33 64 00	24 0831891	1724138
531	28 19 61	23.0434372	1883239	581	33 75 61	24 1039416	1721170
532	28 30 24	23.0651252	1879699	582	33 87 24	24 1246762	1718213
533	28 40 89	23.0867928	1876173	583	33 98 89	24 1453929	1715266
534	28 51 56	23.1084400	1872659	584	34 10 56	24 1660919	1712329
535	28 62 25	23.1300670	1869159	585	34 22 25	24 1867732	1709402
536	28 72 96	23.1516738	1865672	586	34 33 96	24 2074369	1706485
537	28 83 69	23.1732605	1862197	587	34 45 69	24.2280829	1703578
538	28 94 44	23.1948270	1858736	588	34 57 44	24 2487113	1700680
539	29 05 21	23.2163735	1855288	589	34 69 21	24.2693222	1697793
540	29 16 00	23.2379001	1851852	590	34 81 00	24 2899156	1694915
541	29 26 81	23.2594067	1848429	591	34 92 81	24 3104916	1692047
542	29 37 64	23.2808935	1845018	592	35 04 64	24 3310501	1689189
543	29 48 49	23.3023604	1841621	593	35 16 49	24 3515913	1686341
544	29 59 36	23.3238076	1838235	594	35 28 36	24 3721152	1683502
545	29 70 25	23.3452351	1834862	595	35 40 25	24 3926218	1680672
546	29 81 16	23.3666429	1831502	596	35 52 16	24 4131112	1677852
547	29 92 09	23.3880311	1828154	597	35 64 09	24 4335834	1675042
548	30 03 04	23.4093998	1824818	598	35 76 04	24 4540385	1672241
549	30 14 01	23.4307490	1821494	599	35 88 01	24 4744765	1669449
550	30 25 00	23.4520788	1818182	600	36 00 00	24 4918974	1666667

No.	Square	Square Root	Reciprocal .00	No.	Square	Square Root	Reciprocal .00
601	36 12 01	24.5153013	1663894	651	42 38 01	25 5147016	1536098
602	36 24 04	24.5356883	1661130	652	42 51 04	25 5342907	1533742
603	36 36 09	24.5560583	1658375	653	42 64 09	25.5538647	1531394
604	36 48 16	24.5764115	1655629	654	42 77 16	25.5734237	1529052
605	36 60 25	24.5967478	1652893	655	42 90 25	25 5929678	1526718
606	36 72 36	24.6170673	1650165	656	43 03 36	25.6124969	1524390
607	36 84 49	24.6373700	1647446	657	43 16 49	25 6320112	1522070
608	36 96 64	24.6576560	1644737	658	43 29 64	25.6515107	1519757
609	37 08 81	24.6779254	1642036	659	43 42 81	25 6709953	1517451
610	37 21 00	24.6981781	1639344	660	43 56 00	25 6904652	1515152
611	37 33 21	24 7184142	1636661	661	43 69 21	25 7099203	1512859
612	37 45 44	24.7386338	1633987	662	43 82 44	25.7293607	1510574
613	37 57 69	24.7588368	1631321	663	43 95 69	25.7487864	1508296
614	37 69 96	24.7790234	1628664	664	44 08 96	25.7681975	1506024
615	37 82 25	24.7991935	1626016	665	44 22 25	25 7875939	1503759
616	37 94 56	24.8193473	1623377	666	44 35 56	25 8069758	1501502
617	38 06 89	24.8394847	1620746	667	44 48 89	25.8263431	1499250
618	38 19 24	24.8596058	1618123	668	44 62 24	25.8456960	1497006
619	38 31 61	24.8797106	1615509	669	44 75 61	25.8650343	1494768
620	38 44 00	24.8997992	1612903	670	44 89 00	25.8843582	1492537
621	38 56 41	24.9198716	1610306	671	45 02 41	25.9036677	1490313
622	38 68 84	24.9399278	1607717	672	45 15 84	25.9229628	1488095
623	38 81 29	24.9599679	1605136	673	45 29 29	25 9422435	1485884
624	38 93 76	24.9799920	1602564	674	45 42 76	25 9615100	1483680
625	39 06 25	25.0000000	1600000	675	45 56 25	25 9807621	1481481
626	39 18 76	25 0199920	1597444	676	45 69 76	26.0000000	1479290
627	39 31 29	25.0399681	1594896	677	45 83 29	26 0192237	1477105
628	39 43 84	25.0599282	1592357	678	45 96 84	26 0384331	1474926
629	39 56 41	25.0798724	1589825	679	46 10 41	26.0576284	1472754
630	39 69 00	25.0998008	1587302	680	46 24 00	26 0768096	1470588
631	39 81 61	25 1197134	1584786	681	46 37 61	26 0959767	1468429
632	39 94 24	25.1396102	1582278	682	46 51 24	26.1151297	1466276
633	40 06 89	25.1594913	1579779	683	46 64 89	26 1342687	1464129
634	40 19 56	25.1793566	1577287	684	46 78 56	26 1533937	1461988
635	40 32 25	25.1992063	1574803	685	46 92 25	26 1725047	1459854
636	40 44 96	25.2190404	1572327	686	47 05 96	26.1916017	1457726
637	40 57 69	25.2388589	1569859	687	47 19 69	26.2106848	1455604
638	40 70 44	25.2586619	1567398	688	47 33 44	26.2297541	1453488
639	40 83 21	25.2784493	1564945	689	47 47 21	26.2488095	1451379
640	40 96 00	25.2982213	1562500	690	47 61 00	26.2678511	1449275
641	41 08 81	25.3179778	1560062	691	47 74 81	26 2868789	1447178
642	41 21 64	25.3377189	1557632	692	47 88 64	26 3058929	1445087
643	41 34 49	25.3574447	1555210	693	48 02 49	26.3248932	1443001
644	41 47 36	25.3771551	1552795	694	48 16 36	26 3438797	1440922
645	41 60 25	25.3968502	1550388	695	48 30 25	26.3628527	1438849
646	41 73 16	25.4165301	1547988	696	48 44 16	26.3818119	1436782
647	41 86 09	25.4361947	1545595	697	48 58 09	26.4007576	1434720
648	41 99 04	25.4558441	1543210	698	48 72 04	26.4196896	1432665
649	42 12 01	25.4754784	1540832	699	48 86 01	26.4386081	1430615
650	42 25 00	25.4950976	1538462	700	49 00 00	26.4575131	1428571

No.	Square	Square Root	Reciprocal .00	No.	Square	Square Root	Reciprocal .00
701	49 14 01	26.4764046	1426534	751	56 40 01	27.4043792	1331558
702	49 28 04	26.4952826	1424501	752	56 55 04	27.4226184	1329787
703	49 42 09	26.5141472	1422475	753	56 70 09	27.4408455	1328021
704	49 56 16	26.5329983	1420455	754	56 85 16	27.4590604	1326260
705	49 70 25	26.5518361	1418440	755	57 00 25	27.4772633	1324503
706	49 84 36	26.5706605	1416431	756	57 15 36	27.4954542	1322751
707	49 98 49	26.5894716	1414427	757	57 30 49	27.5136330	1321004
708	50 12 64	26.6082694	1412429	758	57 45 64	27.5317998	1319261
709	50 26 81	26.6270539	1410437	759	57 60 81	27.5499546	1317523
710	50 41 00	26 6458252	1408451	760	57 76 00	27.5680975	1315789
711	50 55 21	26 6645833	1406470	761	57 91 21	27.5862284	1314060
712	50 69 44	26.6833281	1404494	762	58 06 44	27.6043475	1312336
713	50 83 69	26.7020598	1402525	763	58 21 69	27.6224546	1310616
714	50 97 96	26.7207784	1400560	764	58 36 96	27.6405499	1308901
715	51 12 25	26.7394839	1398601	765	58 52 25	27.6586334	1307190
716	51 26 56	26.7581763	1396648	766	58 67 56	27.6767050	1305483
717	51 40 89	26 7768557	1394700	767	58 82 89	27.6947648	1303781
718	51 55 24	26.7955220	1392758	768	58 98 24	27.7128129	1302083
719	51 69 61	26 8141754	1390821	769	59 13 61	27.7308492	1300390
720	51 84 00	26.8328157	1388889	770	59 29 00	27.7488739	1298701
721	51 98 41	26.8514432	1386963	771	59 44 41	27.7668868	1297017
722	52 12 84	26.8700577	1385042	772	59 59 84	27.7848880	1295337
723	52 27 29	26.8886593	1383126	773	59 75 29	27.8028775	1293661
724	52 41 76	26.9072481	1381215	774	59 90 76	27.8208555	1291990
725	52 56 25	26.9258240	1379310	775	60 06 25	27.8388218	1290323
726	52 70 76	26.9443872	1377410	776	60 21 76	27.8567766	1288660
727	52 85 29	26.9629375	1375516	777	60 37 29	27.8747197	1287001
728	52 99 84	26.9814751	1373626	778	60 52 84	27.8926514	1285347
729	53 14 41	27 0000000	1371742	779	60 68 41	27.9105715	1283697
730	53 29 00	27.0185122	1369863	780	60 84 00	27.9284801	1282051
731	53 43 61	27.0370117	1367989	781	60 99 61	27.9463772	1280410
732	53 58 24	27.0554985	1366120	782	61 15 24	27.9642629	1278772
733	53 72 89	27.0739727	1364256	783	61 30 89	27.9821372	1277139
734	53 87 56	27.0924344	1362398	784	61 46 56	28.0000000	1275510
735	54 02 25	27.1108834	1360544	785	61 62 25	28.0178515	1273885
736	54 16 96	27.1293199	1358696	786	61 77 96	28.0356915	1272265
737	54 31 69	27.1477439	1356852	787	61 93 69	28.0535203	1270648
738	54 46 44	27.1661554	1355014	788	62 09 44	28.0713377	1269036
739	54 61 21	27.1845544	1353180	789	62 25 21	28.0891438	1267427
740	54 76 00	27.2029410	1351351	790	62 41 00	28.1069386	1265823
741	54 90 81	27.2213152	1349528	791	62 56 81	28.1247222	1264223
742	55 05 64	27.2396769	1347709	792	62 72 64	28.1424946	1262626
743	55 20 49	27.2580263	1345895	793	62 88 49	28.1602557	1261034
744	55 35 36	27.2763634	1344086	794	63 04 36	28.1780056	1259446
745	55 50 25	27.2946881	1342282	795	63 20 25	28.1957444	1257862
746	55 65 16	27.3130006	1340483	796	63 36 16	28.2134720	1256281
747	55 80 09	27.3313007	1338688	797	63 52 09	28.2311884	1254705
748	55 95 04	27.3495887	1336898	798	63 68 04	28.2488938	1253133
749	56 10 01	27.3678644	1335113	799	63 84 01	28.2665881	1251564
750	56 25 00	27.3861279	1333333	800	64 00 00	28.2842712	1250000

No	Square	Square Root	Reciprocal .00	No.	Square	Square Root	Reciprocal .00
801	64 16 01	28·3019434	1248439	851	72 42 01	29.1719043	1175088
802	64 32 04	28 3196045	1246883	852	72 59 04	29.1890390	1173709
803	64 48 09	28.3372546	1245330	853	72 76 09	29.2061637	1172333
804	64 64 16	28.3548938	1243781	854	72 93 16	29 2232784	1170960
805	64 80 25	28.3725219	1242236	855	73 10 25	29 2403830	1169591
806	64 96 36	28.3901391	1240695	856	73 27 36	29.2574777	1168224
807	65 12 49	28.4077454	1239157	857	73 44 49	29 2745623	1166861
808	65 28 64	28.4253408	1237624	858	73 61 64	29.2916370	1165501
809	65 44 81	28.4429253	1236094	859	73 78 81	29.3087018	1164144
810	65 61 00	28.4604989	1234568	860	73 96 00	29 3257566	1162791
811	65 77 21	28.4780617	1233046	861	74 13 21	29.3428015	1161440
812	65 93 44	28.4956137	1231527	862	74 30 44	29.3598365	1160093
813	66 09 69	28 5131549	1230012	863	74 47 69	29.3768616	1158749
814	66 25 96	28.5306852	1228501	864	74 64 96	29 3938769	1157407
815	66 42 25	28.5482048	1226994	865	74 82 25	29.4108823	1156069
816	66 58 56	28.5657137	1225490	866	74 99 56	29.4278779	1154734
817	66 74 89	28.5832119	1223990	867	75 16 89	29.4448637	1153403
818	66 91 24	28.6006993	1222494	868	75 34 24	29.4618397	1152074
819	67 07 61	28.6181760	1221001	869	75 51 61	29.4788059	1150748
820	67 24 00	28.6356421	1219512	870	75 69 00	29.4957624	1149425
821	67 40 41	28.6530976	1218027	871	75 86 41	29.5127091	1148106
822	67 56 84	28 6705424	1216545	872	76 03 84	29.5296461	1146789
823	67 73 29	28.6879766	1215067	873	76 21 29	29.5465734	1145475
824	67 89 76	28.7054002	1213592	874	76 38 76	29.5634910	1144165
825	68 06 25	28.7228132	1212121	875	76 56 25	29.5803989	1142857
826	68 22 76	28.7402157	1210654	876	76 73 76	29.5972972	1141553
827	68 39 29	28.7576077	1209190	877	76 91 29	29.6141858	1140251
828	68 55 84	28.7749891	1207729	878	77 08 84	29.6310648	1138952
829	68 72 41	28.7923601	1206273	879	77 26 41	29.6479342	1137656
830	68 89 00	28.8097206	1204819	880	77 44 00	29.6647939	1136364
831	69 05 61	28.8270706	1203369	881	77 61 61	29.6816442	1135074
832	69 22 24	28.8444102	1201923	882	77 79 24	29.6984848	1133787
833	69 38 89	28.8617394	1200480	883	77 96 89	29.7153159	1132503
834	69 55 56	28.8790582	1199041	884	78 14 56	29.7321375	1131222
835	69 72 25	28 8963666	1197605	885	78 32 25	29.7489496	1129944
836	69 88 96	28 9136646	1196172	886	78 49 96	29.7657521	1128668
837	70 05 69	28.9309523	1194743	887	78 67 69	29.7825452	1127396
838	70 22 44	28 9482297	1193317	888	78 85 44	29.7993289	1126126
839	70 39 21	28.9654967	1191895	889	79 03 21	29.8161030	1124859
840	70 56 00	28.9827535	1190476	890	79 21 00	29.8328678	1123596
841	70 72 81	29.0000000	1189061	891	79 38 81	29.8496231	1122334
842	70 89 64	29.0172363	1187648	892	79 56 64	29.8663690	1121076
843	71 06 49	29.0344623	1186240	893	79 74 49	29.8831056	1119821
844	71 23 36	29.0516781	1184834	894	79 92 36	29.8998328	1118568
845	71 40 25	29.0688837	1183432	895	80 10 25	29.9165506	1117318
846	71 57 16	29.0860791	1182033	896	80 28 16	29.9332591	1116071
847	71 74 09	29.1032644	1180638	897	80 46 09	29.9499583	1114827
848	71 91 04	29.1204396	1179245	898	80 64 04	29.9666481	1113586
849	72 08 01	29.1376046	1177856	899	80 82 01	29.9833287	1112347
850	72 25 00	29.1547595	1176471	900	81 00 00	30.0000000	1111111

No.	Square	Square Root	Reciprocal .00	No.	Square	Square Root	Reciprocal .00
901	81 18 01	30 0166620	1109878	951	90 44 01	30 8382879	1051525
902	81 36 04	30 0333148	1108647	952	90 63 04	30 8544972	1050420
903	81 54 09	30.0499584	1107420	953	90 82 09	30 8706981	1049318
904	81 72 16	30.0665928	1106195	954	91 01 16	30 8868904	1048218
905	81 90 25	30 0832179	1104972	955	91 20 25	30 9030743	1047120
906	82 08 36	30.0998339	1103753	956	91 39 36	30.9192497	1046025
907	82 26 49	30.1164407	1102536	957	91 58 49	30.9354166	1044932
908	82 44 64	30.1330383	1101322	958	91 77 64	30 9515751	1043841
909	82 62 81	30.1496269	1100110	959	91 96 81	30 9677251	1042753
910	82 81 00	30 1662063	1098901	960	92 16 00	30.9838668	1041667
911	82 99 21	30.1827765	1097695	961	92 35 21	31 0000000	1040583
912	83 17 44	30.1993377	1096491	962	92 54 44	31 0161248	1039501
913	83 35 69	30.2158899	1095290	963	92 73 69	31.0322413	1038422
914	83 53 96	30.2324329	1094092	964	92 92 96	31 0483494	1037344
915	83 72 25	30 2489669	1092896	965	93 12 25	31.0644491	1036269
916	83 90 56	30.2654919	1091703	966	93 31 56	31.0805405	1035197
917	84 08 89	30 2820079	1090513	967	93 50 89	31.0966236	1034126
918	84 27 24	30.2985148	1089325	968	93 70 24	31.1126984	1033058
919	84 45 61	30 3150128	1088139	969	93 89 61	31.1287648	1031992
920	84 64 00	30.3315018	1086957	970	94 09 00	31.1448230	1030928
921	84 82 41	30.3479818	1085776	971	94 28 41	31.1608729	1029866
922	85 00 84	30 3644529	1084599	972	94 47 84	31.1769145	1028807
923	85 19 29	30 3809151	1083424	973	94 67 29	31.1929479	1027749
924	85 37 76	30.3973683	1082251	974	94 86 76	31.2089731	1026694
925	85 56 25	30.4138127	1081081	975	95 06 25	31 2249900	1025641
926	85 74 76	30.4302481	1079914	976	95 25 76	31.2409987	1024590
927	85 93 29	30.4466747	1078749	977	95 45 29	31.2569992	1023541
928	86 11 84	30.4630924	1077586	978	95 64 84	31.2729915	1022495
929	86 30 41	30.4795013	1076426	979	95 84 41	31.2889757	1021450
930	86 49 00	30.4959014	1075269	980	96 04 00	31.3049517	1020408
931	86 67 61	30 5122926	1074114	981	96 23 61	31.3209195	1019368
932	86 86 24	30.5286750	1072961	982	96 43 24	31.3368792	1018330
933	87 04 89	30.5450487	1071811	983	96 62 89	31.3528308	1017294
934	87 23 56	30.5614136	1070664	984	96 82 56	31.3687743	1016260
935	87 42 25	30 5777697	1069519	985	97 02 25	31.3847097	1015228
936	87 60 96	30.5941171	1068376	986	97 21 96	31.4006369	1014199
937	87 79 69	30.6104557	1067236	987	97 41 69	31.4165561	1013171
938	87 98 44	30 6267857	1066098	988	97 61 44	31.4324673	1012146
939	88 17 21	30 6431069	1064963	989	97 81 21	31.4483704	1011122
940	88 36 00	30.6594194	1063830	990	98 01 00	31.4642654	1010101
941	88 54 81	30.6757233	1062699	991	98 20 81	31.4801525	1009082
942	88 73 64	30.6920185	1061571	992	98 40 64	31.4960315	1008065
943	88 92 49	30.7083051	1060445	993	98 60 49	31 5119025	1007049
944	89 11 36	30.7245830	1059322	994	98 80 36	31 5277655	1006036
945	89 30 25	30.7408523	1058201	995	99 00 25	31.5436206	1005025
946	89 49 16	30.7571130	1057082	996	99 20 16	31.5594677	1004016
947	89 68 09	30 7733651	1055936	997	99 40 09	31.5753068	1003009
948	89 87 04	30.7896086	1054852	998	99 60 04	31.5911380	1002004
949	90 06 01	30.8058436	1053741	999	99 80 01	31 6069613	1001001
950	90 25 00	30.8220700	1052632	1000	1 00 00 00	31.6227766	1000000

APPENDIX P

Table of Logarithms

N.	0	1	2	3	4	5	6	7	8	9	D.
100	000000	000434	000868	001301	001734	002166	002598	003029	003461	003891	432
1	4321	4751	5181	5609	6038	6466	6894	7321	7748	8174	428
2	8600	9026	9451	9876	010300	010724	011147	011570	011993	012415	424
3	012837	013259	013680	014100	4521	4940	5360	5779	6197	6616	420
4	7033	7451	7868	8284	8700	9116	9532	9947	020361	020775	416
105	021189	021603	022016	022428	022841	023252	023664	024075	4486	4896	412
6	5306	5715	6125	6533	6942	7350	7757	8164	8571	8978	408
7	9384	9789	030195	030600	031004	031408	031812	032216	032619	033021	404
8	033424	033826	4227	4628	5029	5430	5830	6230	6629	7028	400
9	7426	7825	8223	8620	9017	9414	9811	040207	040602	040998	397
110	041393	041787	042182	042576	042969	043362	043755	044148	044540	044932	393
1	5323	5714	6105	6495	6885	7275	7664	8053	8442	8830	390
2	9218	9606	9993	050380	050766	051153	051538	051924	052309	052694	386
3	053078	053463	053846	4230	4613	4996	5378	5760	6142	6524	383
4	6905	7286	7666	8046	8426	8805	9185	9563	9942	060320	379
115	060698	061075	061452	061829	062206	062582	062958	063333	063709	4083	376
6	4458	4832	5206	5580	5953	6326	6699	7071	7443	7815	373
7	8186	8557	8928	9298	9668	070038	070407	070776	071145	071514	370
8	071882	072250	072617	072985	073352	3718	4085	4451	4816	5182	366
9	5547	5912	6276	6640	7004	7368	7731	8094	8457	8819	363
120	079181	079543	079904	080266	080626	080987	081347	081707	082067	082426	360
1	082785	083144	083503	3861	4219	4576	4934	5291	5647	6004	357
2	6360	6716	7071	7426	7781	8136	8490	8845	9198	9552	355
3	9905	090258	090611	090963	091315	091667	092018	092370	092721	093071	352
4	093422	3772	4122	4471	4820	5169	5518	5866	6215	6562	349
125	6910	7257	7604	7951	8298	8644	8990	9335	9681	100026	346
6	100371	100715	101059	101403	101747	102091	102434	102777	103119	3462	343
7	3804	4146	4487	4828	5169	5510	5851	6191	6531	6871	341
8	7210	7549	7888	8227	8565	8903	9241	9579	9916	110253	338
9	110590	110926	111263	111599	111934	112270	112605	112940	113275	3609	335
130	113943	114277	114611	114944	115278	115611	115943	116276	116608	116940	333
1	7271	7603	7934	8265	8595	8926	9256	9586	9915	120245	330
2	120574	120903	121231	121560	121888	122216	122544	122871	123198	3525	328
3	3852	4178	4504	4830	5156	5481	5806	6131	6456	6781	325
4	7105	7429	7753	8076	8399	8722	9045	9368	9690	130012	323
135	130334	130655	130977	131298	131619	131939	132260	132580	132900	3219	321
6	3539	3858	4177	4496	4814	5133	5451	5769	6086	6403	318
7	6721	7037	7354	7671	7987	8303	8618	8934	9249	9564	316
8	9879	140194	140508	140822	141136	141450	141763	142076	142389	142702	314
9	143015	3327	3639	3951	4263	4574	4885	5196	5507	5818	311
140	146128	146438	146748	147058	147367	147676	147985	148294	148603	148911	309
1	9219	9527	9835	150142	150449	150756	151063	151370	151676	151982	307
2	152288	152594	152900	3205	3510	3815	4120	4424	4728	5032	305
3	5336	5640	5943	6246	6549	6852	7154	7457	7759	8061	303
4	8362	8664	8965	9266	9567	9868	160168	160469	160769	161068	301
145	161368	161667	161967	162266	162564	162863	3161	3460	3758	4055	299
6	4353	4650	4947	5244	5541	5838	6134	6430	6726	7022	297
7	7317	7613	7908	8203	8497	8792	9086	9380	9674	9968	295
8	170262	170555	170848	171141	171434	171726	172019	172311	172603	172895	293
9	3186	3478	3769	4060	4351	4641	4932	5222	5512	5802	291
150	176091	176381	176670	176959	177248	177536	177825	178113	178401	178689	289
1	8977	9264	9552	9839	180126	180413	180699	180986	181272	181558	287
2	181844	182129	182415	182700	2985	3270	3555	3839	4123	4407	285
3	4691	4975	5259	5542	5825	6108	6391	6674	6956	7239	283
4	7521	7803	8084	8366	8647	8928	9209	9490	9771	190051	281
155	190332	190612	190892	191171	191451	191730	192010	192289	192567	2846	279
6	3125	3403	3681	3959	4237	4514	4792	5069	5346	5623	278
7	5900	6176	6453	6729	7005	7281	7556	7832	8107	8382	276
8	8657	8932	9206	9481	9755	200029	200303	200577	200850	201124	274
9	201397	201670	201943	202216	202488	2761	3033	3305	3577	3848	272
N.	0	1	2	3	4	5	6	7	8	9	D.

N.	0	1	2	3	4	5	6	7	8	9	D.
160	204120	204391	204663	204934	205204	205475	205746	206016	206286	206556	271
1	6826	7096	7365	7634	7904	8173	8441	8710	8979	9247	269
2	9515	9783	210051	210319	210586	210853	211121	211388	211654	211921	267
3	212188	212454	2720	2986	3252	3518	3783	4049	4314	4579	266
4	4844	5109	5373	5638	5902	6166	6430	6694	6957	7221	264
165	7484	7747	8010	8273	8536	8798	9060	9323	9585	9846	262
6	220108	220370	220631	220892	221153	221414	221675	221936	222196	222456	261
7	2716	2976	3236	3496	3755	4015	4274	4533	4792	5051	259
8	5309	5568	5826	6084	6342	6600	6858	7115	7372	7630	258
9	7887	8144	8400	8657	8913	9170	9426	9682	9938	230193	256
170	230449	230704	230960	231215	231470	231724	231979	232234	232488	232742	255
1	2996	3250	3504	3757	4011	4264	4517	4770	5023	5276	253
2	5528	5781	6033	6285	6537	6789	7041	7292	7544	7795	252
3	8046	8297	8548	8799	9049	9299	9550	9800	240050	240300	250
4	240549	240799	241048	241297	241546	241795	242044	242293	2541	2790	249
175	3038	3286	3534	3782	4030	4277	4525	4772	5019	5266	248
6	5513	5759	6006	6252	6499	6745	6991	7237	7482	7728	246
7	7973	8219	8464	8709	8954	9198	9443	9687	9932	250176	245
8	250420	250664	250908	251151	251395	251638	251881	252125	252368	2610	243
9	2853	3096	3338	3580	3822	4064	4306	4548	4790	5031	242
180	255273	255514	255755	255996	256237	256477	256718	256958	257198	257439	241
1	7679	7918	8158	8398	8637	8877	9116	9355	9594	9833	239
2	260071	260310	260548	260787	261025	261263	261501	261739	261976	262214	238
3	2451	2688	2925	3162	3399	3636	3873	4109	4346	4582	237
4	4818	5054	5290	5525	5761	5996	6232	6467	6702	6937	235
185	7172	7406	7641	7875	8110	8344	8578	8812	9046	9279	234
6	9513	9746	9980	270213	270446	270679	270912	271144	271377	271609	233
7	271842	272074	272306	2538	2770	3001	3233	3464	3696	3927	232
8	4158	4389	4620	4850	5081	5311	5542	5772	6002	6232	230
9	6462	6692	6921	7151	7380	7609	7838	8067	8296	8525	229
190	278754	278982	279211	279439	279667	279895	280123	280351	280578	280806	228
1	281033	281261	281488	281715	281942	282169	2396	2622	2849	3075	227
2	3301	3527	3753	3979	4205	4431	4656	4882	5107	5332	226
3	5557	5782	6007	6232	6456	6681	6905	7130	7354	7578	225
4	7802	8026	8249	8473	8696	8920	9143	9366	9589	9812	223
195	290035	290257	290480	290702	290925	291147	291369	291591	291813	292034	222
6	2256	2478	2699	2920	3141	3363	3584	3804	4025	4246	221
7	4466	4687	4907	5127	5347	5567	5787	6007	6226	6446	220
8	6665	6884	7104	7323	7542	7761	7979	8198	8416	8635	219
9	8853	9071	9289	9507	9725	9943	300161	300378	300595	300813	218
200	301030	301247	301464	301681	301898	302114	302331	302547	302764	302980	217
1	3196	3412	3628	3844	4059	4275	4491	4706	4921	5136	216
2	5351	5566	5781	5996	6211	6425	6639	6854	7068	7282	215
3	7496	7710	7924	8137	8351	8564	8778	8991	9204	9417	213
4	9630	9843	310056	310268	310481	310693	310906	311118	311330	311542	212
205	311754	311966	2177	2389	2600	2812	3023	3234	3445	3656	211
6	3867	4078	4289	4499	4710	4920	5130	5340	5551	5760	210
7	5970	6180	6390	6599	6809	7018	7227	7436	7646	7854	209
8	8063	8272	8481	8689	8898	9106	9314	9522	9730	9938	208
9	320146	320354	320562	320769	320977	321184	321391	321598	321805	322012	207
210	322219	322426	322633	322839	323046	323252	323458	323665	323871	324077	206
1	4282	4488	4694	4899	5105	5310	5516	5721	5926	6131	205
2	6336	6541	6745	6950	7155	7359	7563	7767	7972	8176	204
3	8380	8583	8787	8991	9194	9398	9601	9805	330008	330211	203
4	330414	330617	330819	331022	331225	331427	331630	331832	2034	2236	202
215	2438	2640	2842	3044	3246	3447	3649	3850	4051	4253	202
6	4454	4655	4856	5057	5257	5458	5658	5859	6059	6260	201
7	6460	6660	6860	7060	7260	7459	7659	7858	8058	8257	200
8	8456	8656	8855	9054	9253	9451	9650	9849	340047	340246	199
9	340444	340642	340841	341039	341237	341435	341632	341830	2028	2225	198
N.	0	1	2	3	4	5	6	7	8	9	D.

N.	0	1	2	3	4	5	6	7	8	9	D.
220	342423	342620	342817	343014	343212	343409	343606	343802	343999	344196	197
1	4392	4589	4785	4981	5178	5374	5570	5766	5962	6157	196
2	6353	6549	6744	6939	7135	7330	7525	7720	7915	8110	195
3	8305	8500	8694	8889	9083	9278	9472	9666	9860	350054	194
4	350248	350442	350636	350829	351023	351216	351410	351603	351796	1989	193
225	2183	2375	2568	2761	2954	3147	3339	3532	3724	3916	193
6	4108	4301	4493	4685	4876	5068	5260	5452	5643	5834	192
7	6026	6217	6408	6599	6790	6981	7172	7363	7554	7744	191
8	7935	8125	8316	8506	8696	8886	9076	9266	9456	9646	190
9	9835	360025	360215	360404	360593	360783	360972	361161	361350	361539	189
230	361728	361917	362105	362294	362482	362671	362859	363048	363236	363424	188
1	3612	3800	3988	4176	4363	4551	4739	4926	5113	5301	188
2	5488	5675	5862	6049	6236	6423	6610	6796	6983	7169	187
3	7356	7542	7729	7915	8101	8287	8473	8659	8845	9030	186
4	9216	9401	9587	9772	9958	370143	370328	370513	370698	370883	185
235	371068	371253	371437	371622	371806	1991	2175	2360	2544	2728	184
6	2912	3096	3280	3464	3647	3831	4015	4198	4382	4565	184
7	4748	4932	5115	5298	5481	5664	5846	6029	6212	6394	183
8	6577	6759	6942	7124	7306	7488	7670	7852	8034	8216	182
9	8398	8580	8761	8943	9124	9306	9487	9668	9849	380030	181
240	380211	380392	380573	380754	380934	381115	381296	381476	381656	381837	181
1	2017	2197	2377	2557	2737	2917	3097	3277	3456	3636	180
2	3815	3995	4174	4353	4533	4712	4891	5070	5249	5428	179
3	5606	5785	5964	6142	6321	6499	6677	6856	7034	7212	178
4	7390	7568	7746	7923	8101	8279	8456	8634	8811	8989	178
245	9166	9343	9520	9698	9875	390051	390228	390405	390582	390759	177
6	390935	391112	391288	391464	391641	1817	1993	2169	2345	2521	176
7	2697	2873	3048	3224	3400	3575	3751	3926	4101	4277	176
8	4452	4627	4802	4977	5152	5326	5501	5676	5850	6025	175
9	6199	6374	6548	6722	6896	7071	7245	7419	7592	7766	174
250	397940	398114	398287	398461	398634	398808	398981	399154	399328	399501	173
1	9674	9847	400020	400192	400365	400538	400711	400883	401056	401228	173
2	401401	401573	1745	1917	2089	2261	2433	2605	2777	2949	172
3	3121	3292	3464	3635	3807	3978	4149	4320	4492	4663	171
4	4834	5005	5176	5346	5517	5688	5858	6029	6199	6370	171
255	6540	6710	6881	7051	7221	7391	7561	7731	7901	8070	170
6	8240	8410	8579	8749	8918	9087	9257	9426	9595	9764	169
7	9933	410102	410271	410440	410609	410777	410946	411114	411283	411451	169
8	411620	1788	1956	2124	2293	2461	2629	2796	2964	3132	168
9	3300	3467	3635	3803	3970	4137	4305	4472	4639	4806	167
260	414973	415140	415307	415474	415641	415808	415974	416141	416308	416474	167
1	6641	6807	6973	7139	7306	7472	7638	7804	7970	8135	166
2	8301	8467	8633	8798	8964	9129	9295	9460	9625	9791	165
3	9956	420121	420286	420451	420616	420781	420945	421110	421275	421439	165
4	421604	1768	1933	2097	2261	2426	2590	2754	2918	3082	164
265	3246	3410	3574	3737	3901	4065	4228	4392	4555	4718	164
6	4882	5045	5208	5371	5534	5697	5860	6023	6186	6349	163
7	6511	6674	6836	6999	7161	7324	7486	7648	7811	7973	162
8	8135	8297	8459	8621	8783	8944	9106	9268	9429	9591	162
9	9752	9914	430075	430236	430398	430559	430720	430881	431042	431203	161
270	431364	431525	431685	431846	432007	432167	432328	432488	432649	432809	161
1	2969	3130	3290	3450	3610	3770	3930	4090	4249	4409	160
2	4569	4729	4888	5048	5207	5367	5526	5685	5844	6004	159
3	6163	6322	6481	6640	6799	6957	7116	7275	7433	7592	159
4	7751	7909	8067	8226	8384	8542	8701	8859	9017	9175	158
275	9333	9491	9648	9806	9964	440122	440279	440437	440594	440752	158
6	440909	441066	441224	441381	441538	1695	1852	2009	2166	2323	157
7	2480	2637	2793	2950	3106	3263	3419	3576	3732	3889	157
8	4045	4201	4357	4513	4669	4825	4981	5137	5293	5449	156
9	5604	5760	5915	6071	6226	6382	6537	6692	6848	7003	155
N.	0	1	2	3	4	5	6	7	8	9	D.

N.	0	1	2	3	4	5	6	7	8	9	D.
280	447158	447313	447468	447623	447778	447933	448088	448242	448397	448552	155
1	8706	8861	9015	9170	9324	9478	9633	9787	9941	450095	154
2	450249	450403	450557	450711	450865	451018	451172	451326	451479	1633	154
3	1786	1940	2093	2247	2400	2553	2706	2859	3012	3165	153
4	3318	3471	3624	3777	3930	4082	4235	4387	4540	4692	153
285	4845	4997	5150	5302	5454	5606	5758	5910	6062	6214	152
6	6366	6518	6670	6821	6973	7125	7276	7428	7579	7731	152
7	7882	8033	8184	8336	8487	8638	8789	8940	9091	9242	151
8	9392	9543	9694	9845	9995	460146	460296	460447	460597	460748	151
9	460898	461048	461198	461348	461499	1649	1799	1948	2098	2248	150
290	462398	462548	462697	462847	462997	463146	463296	463445	463594	463744	150
1	3893	4042	4191	4340	4490	4539	4788	4936	5085	5234	149
2	5383	5532	5680	5829	5977	6126	6274	6423	6571	6719	149
3	6868	7016	7164	7312	7460	7608	7756	7904	8052	8200	148
4	8347	8495	8643	8790	8938	9085	9233	9380	9527	9675	148
295	9822	9969	470116	470263	470410	470557	470704	470851	470998	471145	147
6	471292	471438	1585	1732	1878	2025	2171	2318	2464	2610	146
7	2756	2903	3049	3195	3341	3487	3633	3779	3925	4071	146
8	4216	4362	4508	4653	4799	4944	5090	5235	5381	5526	146
9	5671	5816	5962	6107	6252	6397	6542	6687	6832	6976	145
300	477121	477266	477411	477555	477700	477844	477989	478133	478278	478422	145
1	8566	8711	8855	8999	9143	9287	9431	9575	9719	9863	144
2	480007	480151	480294	480438	480582	480725	480869	481012	481156	481299	144
3	1443	1586	1729	1872	2016	2159	2302	2445	2588	2731	143
4	2874	3016	3159	3302	3445	3587	3730	3872	4015	4157	143
305	4300	4442	4585	4727	4869	5011	5153	5295	5437	5579	142
6	5721	5863	6005	6147	6239	6430	6572	6714	6855	6997	142
7	7138	7280	7421	7563	7704	7845	7986	8127	8269	8410	141
8	8551	8692	8833	8974	9114	9255	9396	9537	9677	9818	141
9	9958	490099	490239	490380	490520	490661	490801	490941	491081	491222	140
310	491362	491502	491642	491782	491922	492062	492201	492341	492481	492621	140
1	2760	2900	3040	3179	3319	3458	3597	3737	3876	4015	139
2	4155	4294	4433	4572	4711	4850	4989	5128	5267	5406	139
3	5544	5683	5822	5960	6099	6238	6376	6515	6653	6791	139
4	6930	7068	7206	7344	7483	7621	7759	7897	8035	8173	138
315	8311	8448	8586	8724	8862	8999	9137	9275	9412	9550	138
6	9687	9824	9962	500099	500236	500374	500511	500648	500785	500922	137
7	501059	501196	501333	1470	1607	1744	1880	2017	2154	2291	137
8	2427	2564	2700	2837	2973	3109	3246	3382	3518	3655	136
9	3791	3927	4063	4199	4335	4471	4607	4743	4878	5014	136
320	505150	505286	505421	505557	505693	505828	505964	506099	506234	506370	136
1	6505	6640	6776	6911	7046	7181	7316	7451	7586	7721	135
2	7856	7991	8126	8260	8395	8530	8664	8799	8934	9068	135
3	9203	9337	9471	9606	9740	9874	510009	510143	510277	510411	134
4	510545	510679	510813	510947	511081	511215	1349	1482	1616	1750	134
325	1883	2017	2151	2284	2418	2551	2684	2818	2951	3084	133
6	3218	3351	3484	3617	3750	3883	4016	4149	4282	4415	133
7	4548	4681	4813	4946	5079	5211	5344	5476	5609	5741	133
8	5874	6006	6139	6271	6403	6535	6668	6800	6932	7064	132
9	7196	7328	7460	7592	7724	7855	7987	8119	8251	8382	132
330	518514	518646	518777	518909	519040	519171	519303	519434	519566	519697	131
1	9828	9959	520090	520221	520353	520484	520615	520745	520876	521007	131
2	521138	521269	1400	1530	1661	1792	1922	2053	2183	2314	131
3	2444	2575	2705	2835	2966	3096	3226	3356	3486	3616	130
4	3746	3876	4006	4136	4266	4396	4526	4656	4785	4915	130
335	5045	5174	5304	5434	5563	5693	5822	5951	6081	6210	129
6	6339	6469	6598	6727	6856	6985	7114	7243	7372	7501	129
7	7630	7759	7888	8016	8145	8274	8402	8531	8660	8788	129
8	8917	9045	9174	9302	9430	9559	9687	9815	9943	530072	128
9	530200	530328	530456	530584	530712	530840	530968	531096	531223	1351	128
N.	0	1	2	3	4	5	6	7	8	9	D.

N.	0	1	2	3	4	5	6	7	8	9	D.
340	531479	531607	531734	531862	531990	532117	532245	532372	532500	532627	128
1	2754	2882	3009	3136	3264	3391	3518	3645	3772	3899	127
2	4026	4153	4280	4407	4534	4661	4787	4914	5041	5167	127
3	5294	5421	5547	5674	5800	5927	6053	6180	6306	6432	126
4	6558	6685	6811	6937	7063	7189	7315	7441	7567	7693	126
345	7819	7945	8071	8197	8322	8448	8574	8699	8825	8951	126
6	9076	9202	9327	9452	9578	9703	9829	9954	540079	540204	125
7	540329	540455	540580	540705	540830	540955	541080	541205	1330	1454	125
8	1579	1704	1829	1953	2078	2203	2327	2452	2576	2701	125
9	2825	2950	3074	3199	3323	3447	3571	3696	3820	3944	124
350	544068	544192	544316	544440	544564	544688	544812	544936	545060	545183	124
1	5307	5431	5555	5678	5802	5925	6049	6172	6296	6419	124
2	6543	6666	6789	6913	7036	7159	7282	7405	7529	7652	123
3	7775	7898	8021	8144	8267	8389	8512	8635	8758	8881	123
4	9003	9126	9249	9371	9494	9616	9739	9861	9984	550106	123
355	550228	550351	550473	550595	550717	550840	550962	551084	551206	1328	122
6	1450	1572	1694	1816	1938	2060	2181	2303	2425	2547	122
7	2668	2790	2911	3033	3155	3276	3398	3519	3640	3762	121
8	3883	4004	4126	4247	4368	4489	4610	4731	4852	4973	121
9	5094	5215	5336	5457	5578	5699	5820	5940	6061	6182	121
360	556303	556423	556544	556664	556785	556905	557026	557146	557267	557387	120
1	7507	7627	7748	7868	7988	8108	8228	8349	8469	8589	120
2	8709	8829	8948	9068	9188	9308	9428	9548	9667	9787	120
3	9907	560026	560146	560265	560385	560504	560624	560743	560863	560982	119
4	561101	1221	1340	1459	1578	1698	1817	1936	2055	2174	119
365	2293	2412	2531	2650	2769	2887	3006	3125	3244	3362	119
6	3481	3600	3718	3837	3955	4074	4192	4311	4429	4548	119
7	4666	4784	4903	5021	5139	5257	5376	5494	5612	5730	118
8	5848	5966	6084	6202	6320	6437	6555	6673	6791	6909	118
9	7026	7144	7262	7379	7497	7614	7732	7849	7967	8084	118
370	568202	568319	568436	568554	568671	568788	568905	569023	569140	569257	117
1	9374	9491	9608	9725	9842	9959	570076	570193	570309	570426	117
2	570543	570660	570776	570893	571010	571126	1243	1359	1476	1592	117
3	1709	1825	1942	2058	2174	2291	2407	2523	2639	2755	116
4	2872	2988	3104	3220	3336	3452	3568	3684	3800	3915	116
375	4031	4147	4263	4379	4494	4610	4726	4841	4957	5072	116
6	5188	5303	5419	5534	5650	5765	5880	5996	6111	6226	115
7	6341	6457	6572	6687	6802	6917	7032	7147	7262	7377	115
8	7492	7607	7722	7836	7951	8066	8181	8295	8410	8525	115
9	8639	8754	8868	8983	9097	9212	9326	9441	9555	9669	114
380	579784	579898	580012	580126	580241	580355	580469	580583	580697	580811	114
1	580925	581039	1153	1267	1381	1495	1608	1722	1836	1950	114
2	2063	2177	2291	2404	2518	2631	2745	2858	2972	3085	114
3	3199	3312	3426	3539	3652	3765	3879	3992	4105	4218	113
4	4331	4444	4557	4670	4783	4896	5009	5122	5235	5348	113
385	5461	5574	5686	5799	5912	6024	6137	6250	6362	6475	113
6	6587	6700	6812	6925	7037	7149	7262	7374	7486	7599	112
7	7711	7823	7935	8047	8160	8272	8384	8496	8608	8720	112
8	8832	8944	9056	9167	9279	9391	9503	9615	9726	9838	112
9	9950	590061	590173	590284	590396	590507	590619	590730	590842	590953	112
390	591065	591176	591287	591399	591510	591621	591732	591843	591955	592066	111
1	2177	2288	2399	2510	2621	2732	2843	2954	3064	3175	111
2	3286	3397	3508	3618	3729	3840	3950	4061	4171	4282	111
3	4393	4503	4614	4724	4834	4945	5055	5165	5276	5386	110
4	5496	5606	5717	5827	5937	6047	6157	6267	6377	6487	110
395	6597	6707	6817	6927	7037	7146	7256	7366	7476	7586	110
6	7695	7805	7914	8024	8134	8243	8353	8462	8572	8681	110
7	8791	8900	9009	9119	9228	9337	9446	9556	9665	9774	109
8	9883	9992	600101	600210	600319	600428	600537	600646	600755	600864	109
9	600973	601082	1191	1299	1408	1517	1625	1734	1843	1951	109
N.	0	1	2	3	4	5	6	7	8	9	D.

N.	0	1	2	3	4	5	6	7	8	9	D.
400	602060	602169	602277	602386	602494	602603	602711	602819	602928	603036	108
1	3144	3253	3361	3469	3577	3686	3794	3902	4010	4118	108
2	4226	4334	4442	4550	4658	4766	4874	4982	5089	5197	108
3	5305	5413	5521	5628	5736	5844	5951	6059	6166	6274	108
4	6381	6489	6596	6704	6811	6919	7026	7133	7241	7348	107
405	7455	7562	7669	7777	7884	7991	8098	8205	8312	8419	107
6	8526	8633	8740	8847	8954	9061	9167	9274	9381	9488	107
7	9594	9701	9808	9914	610021	610128	610234	610341	610447	610554	107
8	610660	610767	610873	610979	1086	1192	1298	1405	1511	1617	106
9	1723	1829	1936	2042	2148	2254	2360	2466	2572	2678	106
410	612784	612890	612996	613102	613207	613313	613419	613525	613630	613736	106
1	3842	3947	4053	4159	4264	4370	4475	4581	4686	4792	106
2	4897	5003	5108	5213	5319	5424	5529	5634	5740	5845	105
3	5950	6055	6160	6265	6370	6476	6581	6686	6790	6895	105
4	7000	7105	7210	7315	7420	7525	7629	7734	7839	7943	105
415	8048	8153	8257	8362	8466	8571	8676	8780	8884	8989	105
6	9093	9198	9302	9406	9511	9615	9719	9824	9928	620032	104
7	620136	620240	620344	620448	620552	620656	620760	620864	620968	1072	104
8	1176	1280	1384	1488	1592	1695	1799	1903	2007	2110	104
9	2214	2318	2421	2525	2628	2732	2835	2939	3042	3146	104
420	623249	623353	623456	623559	623663	623766	623869	623973	624076	624179	103
1	4282	4385	4488	4591	4695	4798	4901	5004	5107	5210	103
2	5312	5415	5518	5621	5724	5827	5929	6032	6135	6238	103
3	6340	6443	6546	6648	6751	6853	6956	7058	7161	7263	103
4	7366	7468	7571	7673	7775	7878	7980	8082	8185	8287	102
425	8389	8491	8593	8695	8797	8900	9002	9104	9206	9308	102
6	9410	9512	9613	9715	9817	9919	630021	630123	630224	630326	102
7	630428	630530	630631	630733	630835	630936	1038	1139	1241	1342	102
8	1444	1545	1647	1748	1849	1951	2052	2153	2255	2356	101
9	2457	2559	2660	2761	2862	2963	3064	3165	3266	3367	101
430	633468	633569	633670	633771	633872	633973	634074	634175	634276	634376	101
1	4477	4578	4679	4779	4880	4981	5081	5182	5283	5383	101
2	5484	5584	5685	5785	5886	5986	6087	6187	6287	6388	100
3	6488	6588	6688	6789	6889	6989	7089	7189	7290	7390	100
4	7490	7590	7690	7790	7890	7990	8090	8190	8290	8389	100
435	8489	8589	8689	8789	8888	8988	9088	9188	9287	9387	100
6	9486	9586	9686	9785	9885	9984	640084	640183	640283	640382	99
7	640481	640581	640680	640779	640879	640978	1077	1177	1276	1375	99
8	1474	1573	1672	1771	1871	1970	2069	2168	2267	2366	99
9	2465	2563	2662	2761	2860	2959	3058	3156	3255	3354	99
440	643453	643551	643650	643749	643847	643946	644044	644143	644242	644340	98
1	4439	4537	4636	4734	4832	4931	5029	5127	5226	5324	98
2	5422	5521	5619	5717	5815	5913	6011	6110	6208	6306	98
3	6404	6502	6600	6698	6796	6894	6992	7089	7187	7285	98
4	7383	7481	7579	7676	7774	7872	7969	8067	8165	8262	98
445	8360	8458	8555	8653	8750	8848	8945	9043	9140	9237	97
6	9335	9432	9530	9627	9724	9821	9919	650016	650113	650210	97
7	650308	650405	650502	650599	650696	650793	650890	0987	1084	1181	97
8	1278	1375	1472	1569	1666	1762	1859	1956	2053	2150	97
9	2246	2343	2440	2536	2633	2730	2826	2923	3019	3116	97
450	653213	653309	653405	653502	653598	653695	653791	653888	653984	654080	96
1	4177	4273	4369	4465	4562	4658	4754	4850	4946	5042	96
2	5138	5235	5331	5427	5523	5619	5715	5810	5906	6002	96
3	6098	6194	6290	6386	6482	6577	6673	6769	6864	6960	96
4	7056	7152	7247	7343	7438	7534	7629	7725	7820	7916	96
455	8011	8107	8202	8298	8393	8488	8584	8679	8774	8870	95
6	8965	9060	9155	9250	9346	9441	9536	9631	9726	9821	95
7	9916	660011	660106	660201	660296	660391	660486	660581	660676	660771	95
8	660865	0960	1055	1150	1245	1339	1434	1529	1623	1718	95
9	1813	1907	2002	2096	2191	2286	2380	2475	2569	2663	95
N.	0	1	2	3	4	5	6	7	8	9	D.

N.	0	1	2	3	4	5	6	7	8	9	D.
460	662758	662852	662947	663041	663135	663230	663324	663418	663512	663607	94
1	3701	3795	3889	3983	4078	4172	4266	4360	4454	4548	94
2	4642	4736	4830	4924	5018	5112	5206	5299	5393	5487	94
3	5581	5675	5769	5862	5956	6050	6143	6237	6331	6424	94
4	6518	6612	6705	6799	6892	6986	7079	7173	7266	7360	94
465	7453	7546	7640	7733	7826	7920	8013	8106	8199	8293	93
6	8386	8479	8572	8665	8759	8852	8945	9038	9131	9224	93
7	9317	9410	9503	9595	9689	9782	9875	9967	670060	670153	93
8	670246	670339	670431	670524	670617	670710	670802	670895	0988	1080	93
9	1173	1265	1358	1451	1543	1636	1728	1821	1913	2005	93
470	672098	672190	672283	672375	672467	672560	672652	672744	672836	672929	92
1	3021	3113	3205	3297	3390	3482	3574	3666	3758	3850	92
2	3942	4034	4126	4218	4310	4402	4494	4586	4677	4769	92
3	4861	4953	5045	5137	5228	5320	5412	5503	5595	5687	92
4	5778	5870	5962	6053	6145	6236	6328	6419	6511	6602	92
475	6694	6785	6876	6968	7059	7151	7242	7333	7424	7516	91
6	7607	7698	7789	7881	7972	8063	8154	8245	8336	8427	91
7	8518	8609	8700	8791	8882	8973	9064	9155	9246	9337	91
8	9428	9519	9610	9700	9791	9882	9973	680063	680154	680245	91
9	680336	680426	680517	680607	680698	680789	680879	0970	1060	1151	91
480	681241	681332	681422	681513	681603	681693	681784	681874	681964	682055	90
1	2145	2235	2326	2416	2506	2596	2686	2777	2867	2957	90
2	3047	3137	3227	3317	3407	3497	3587	3677	3767	3857	90
3	3947	4037	4127	4217	4307	4396	4486	4576	4666	4756	90
4	4845	4935	5025	5114	5204	5294	5383	5473	5563	5652	90
485	5742	5831	5921	6010	6100	6189	6279	6368	6458	6547	89
6	6636	6726	6815	6904	6994	7083	7172	7261	7351	7440	89
7	7529	7618	7707	7796	7886	7975	8064	8153	8242	8331	89
8	8420	8509	8598	8687	8776	8865	8953	9042	9131	9220	89
9	9309	9398	9486	9575	9664	9753	9841	9930	690019	690107	89
490	690196	690285	690373	690462	690550	690639	690728	690816	690905	690993	89
1	1081	1170	1258	1347	1435	1524	1612	1700	1789	1877	88
2	1965	2053	2142	2230	2318	2406	2494	2583	2671	2759	88
3	2847	2935	3023	3111	3199	3287	3375	3463	3551	3639	88
4	3727	3815	3903	3991	4078	4166	4254	4342	4430	4517	88
495	4605	4693	4781	4868	4956	5044	5131	5219	5307	5394	88
6	5482	5569	5657	5744	5832	5919	6007	6094	6182	6269	87
7	6356	6444	6531	6618	6706	6793	6880	6968	7055	7142	87
8	7229	7317	7404	7491	7578	7665	7752	7839	7926	8014	87
9	8101	8188	8275	8362	8449	8535	8622	8709	8796	8883	87
500	698970	699057	699144	699231	699317	699404	699491	699578	699664	699751	87
1	9838	9924	700011	700098	700184	700271	700358	700444	700531	700617	87
2	700704	700790	0877	0963	1050	1136	1222	1309	1395	1482	86
3	1568	1654	1741	1827	1913	1999	2086	2172	2258	2344	86
4	2431	2517	2603	2689	2775	2861	2947	3033	3119	3205	86
505	3291	3377	3463	3549	3635	3721	3807	3893	3979	4065	86
6	4151	4236	4322	4408	4494	4579	4665	4751	4837	4922	86
7	5008	5094	5179	5265	5350	5436	5522	5607	5693	5778	86
8	5864	5949	6035	6120	6206	6291	6376	6462	6547	6632	85
9	6718	6803	6888	6974	7059	7144	7229	7315	7400	7485	85
510	707570	707655	707740	707826	707911	707996	708081	708166	708251	708336	85
1	8421	8506	8591	8676	8761	8846	8931	9015	9100	9185	85
2	9270	9355	9440	9524	9609	9694	9779	9863	9948	710033	85
3	710117	710202	710287	710371	710456	710540	710625	710710	710794	0879	85
4	0963	1048	1132	1217	1301	1385	1470	1554	1639	1723	84
515	1807	1892	1976	2060	2144	2229	2313	2397	2481	2566	84
6	2650	2734	2818	2902	2986	3070	3154	3238	3323	3407	84
7	3491	3575	3659	3742	3826	3910	3994	4078	4162	4246	84
8	4330	4414	4497	4581	4665	4749	4833	4916	5000	5084	84
9	5167	5251	5335	5418	5502	5586	5669	5753	5836	5920	84
N.	0	1	2	3	4	5	6	7	8	9	D.

N.	0	1	2	3	4	5	6	7	8	9	D
520	716003	716087	716170	716254	716337	716421	716504	716588	716671	716754	83
1	6838	6921	7004	7088	7171	7254	7338	7421	7504	7587	83
2	7671	7754	7837	7920	8003	8086	8169	8253	8336	8419	83
3	8502	8585	8668	8751	8834	8917	9000	9083	9165	9248	83
4	9331	9414	9497	9580	9663	9745	9828	9911	9994	720077	83
525	720159	720242	720325	720407	720490	720573	720655	720738	720821	0903	83
6	0986	1068	1151	1233	1316	1398	1481	1563	1646	1728	82
7	1811	1893	1975	2058	2140	2222	2305	2387	2469	2552	82
8	2634	2716	2798	2881	2963	3045	3127	3209	3291	3374	82
9	3456	3538	3620	3702	3784	3866	3948	4030	4112	4194	82
530	724276	724358	724440	724522	724604	724685	724767	724849	724931	725013	82
1	5095	5176	5258	5340	5422	5503	5585	5667	5748	5830	82
2	5912	5993	6075	6156	6238	6320	6401	6483	6564	6646	82
3	6727	6809	6890	6972	7053	7134	7216	7297	7379	7460	81
4	7541	7623	7704	7785	7866	7948	8029	8110	8191	8273	81
535	8354	8435	8516	8597	8678	8759	8841	8922	9003	9084	81
6	9165	9246	9327	9408	9489	9570	9651	9732	9813	9893	81
7	9974	730055	730136	730217	730298	730378	730459	730540	730621	730702	81
8	730782	0863	0944	1024	1105	1186	1266	1347	1428	1508	81
9	1589	1669	1750	1830	1911	1991	2072	2152	2233	2313	81
540	732394	732474	732555	732635	732715	732796	732876	732956	733037	733117	80
1	3197	3278	3358	3438	3518	3598	3679	3759	3839	3919	80
2	3999	4079	4160	4240	4320	4400	4480	4560	4640	4720	80
3	4800	4880	4960	5040	5120	5200	5279	5359	5439	5519	80
4	5599	5679	5759	5838	5918	5998	6078	6157	6237	6317	80
545	6397	6476	6556	6635	6715	6795	6874	6954	7034	7113	80
6	7193	7272	7352	7431	7511	7590	7670	7749	7829	7908	79
7	7987	8067	8146	8225	8305	8384	8463	8543	8622	8701	79
8	8781	8860	8939	9018	9097	9177	9256	9335	9414	9493	79
9	9572	9651	9731	9810	9889	9968	740047	740126	740205	740284	79
550	740363	740442	740521	740600	740678	740757	740836	740915	740994	741073	79
1	1152	1230	1309	1388	1467	1546	1624	1703	1782	1860	79
2	1939	2018	2096	2175	2254	2332	2411	2489	2568	2647	79
3	2725	2804	2882	2961	3039	3118	3196	3275	3353	3431	78
4	3510	3588	3667	3745	3823	3902	3980	4058	4136	4215	78
555	4293	4371	4449	4528	4606	4684	4762	4840	4919	4997	78
6	5075	5153	5231	5309	5387	5465	5543	5621	5699	5777	78
7	5855	5933	6011	6089	6167	6245	6323	6401	6479	6556	78
8	6634	6712	6790	6868	6945	7023	7101	7179	7256	7334	78
9	7412	7489	7567	7645	7722	7800	7878	7955	8033	8110	78
560	748188	748266	748343	748421	748498	748576	748653	748731	748808	748885	77
1	8963	9040	9118	9195	9272	9350	9427	9504	9582	9659	77
2	9736	9814	9891	9968	750045	750123	750200	750277	750354	750431	77
3	750508	750586	750663	750740	0817	0894	0971	1048	1125	1202	77
4	1279	1356	1433	1510	1587	1664	1741	1818	1895	1972	77
565	2048	2125	2202	2279	2356	2433	2509	2586	2663	2740	77
6	2816	2893	2970	3047	3123	3200	3277	3353	3430	3506	77
7	3583	3660	3736	3813	3889	3966	4042	4119	4195	4272	77
8	4348	4425	4501	4578	4654	4730	4807	4883	4960	5036	76
9	5112	5189	5265	5341	5417	5494	5570	5646	5722	5799	76
570	755875	755951	756027	756103	756180	756256	756332	756408	756484	756560	76
1	6636	6712	6788	6864	6940	7016	7092	7168	7244	7320	76
2	7396	7472	7548	7624	7700	7775	7851	7927	8003	8079	76
3	8155	8230	8306	8382	8458	8533	8609	8685	8761	8836	76
4	8912	8988	9063	9139	9214	9290	9366	9441	9517	9592	76
575	9668	9743	9819	9894	9970	760045	760121	760196	760272	760347	75
6	760422	760498	760573	760649	760724	0799	0875	0950	1025	1101	75
7	1176	1251	1326	1402	1477	1552	1627	1702	1778	1853	75
8	1928	2003	2078	2153	2228	2303	2378	2453	2529	2604	75
9	2679	2754	2829	2904	2978	3053	3128	3203	3278	3353	75
N.	0	1	2	3	4	5	6	7	8	9	D.

N.	0	1	2	3	4	5	6	7	8	9	D.
580	763428	763503	763578	763653	763727	763802	763877	763952	764027	764101	75
1	4176	4251	4326	4400	4475	4550	4624	4699	4774	4848	75
2	4923	4998	5072	5147	5221	5296	5370	5445	5520	5594	75
3	5669	5743	5818	5892	5966	6041	6115	6190	6264	6338	74
4	6413	6487	6562	6636	6710	6785	6859	6933	7007	7082	74
585	7156	7230	7304	7379	7453	7527	7601	7675	7749	7823	74
6	7898	7972	8046	8120	8194	8268	8342	8416	8490	8564	74
7	8638	8712	8786	8860	8934	9008	9082	9156	9230	9303	74
8	9377	9451	9525	9599	9673	9746	9820	9894	9968	770042	74
9	770115	770189	770263	770336	770410	770484	770557	770631	770705	0778	74
590	770852	770926	770999	771073	771146	771220	771293	771367	771440	771514	74
1	1587	1661	1734	1808	1881	1955	2028	2102	2175	2248	73
2	2322	2395	2468	2542	2615	2688	[2762	2835	2908	2981	73
3	3055	3128	3201	3274	3348	3421	3494	3567	3640	3713	73
4	3786	3860	3933	4006	4079	4152	4225	4298	4371	4444	73
595	4517	4590	4663	4736	4809	4882	4955	5028	5100	5173	73
6	5246	5319	5392	5465	5538	5610	5683	5756	5829	5902	73
7	5974	6047	6120	6193	6265	6338	6411	6483	6556	6629	73
8	6701	6774	6846	6919	6992	7064	7137	7209	7282	7354	73
9	7427	7499	7572	7644	7717	7789	7862	7934	8006	8079	72
600	778151	778224	778296	778368	778441	778513	778585	778658	778730	778802	72
1	8874	8947	9019	9091	9163	9236	9308	9380	9452	9524	72
2	9596	9669	9741	9813	9885	9957	780029	780101	780173	780245	72
3	780317	780389	780461	780533	780605	780677	0749	0821	0893	0965	72
4	1037	1109	1181	1253	1324	1396	1468	1540	1612	1684	72
605	1755	1827	1899	1971	2042	2114	2186	2258	2329	2401	72
6	2473	2544	2616	2688	2759	2831	2902	2974	3046	3117	72
7	3189	3260	3332	3403	3475	3546	3618	3689	3761	3832	71
8	3904	3975	4046	4118	4189	4261	4332	4403	4475	4546	71
9	4617	4689	4760	4831	4902	4974	5045	5116	5187	5259	71
610	785330	785401	785472	785543	785615	785686	785757	785828	785899	785970	71
1	6041	6112	6183	6254	6325	6396	6467	6538	6609	6680	71
2	6751	6822	6893	6964	7035	7106	7177	7248	7319	7390	71
3	7460	7531	7602	7673	7744	7815	7885	7956	8027	8098	71
4	8168	8239	8310	8381	8451	8522	8593	8663	8734	8804	71
615	8875	8946	9016	9087	9157	9228	9299	9369	9440	9510	71
6	9581	9651	9722	9792	9863	9933	790004	790074	790144	790215	70
7	790285	790356	790426	790496	790567	790637	0707	0778	0848	0918	70
8	0988	1059	1129	1199	1269	1340	1410	1480	1550	1620	70
9	1691	1761	1831	1901	1971	2041	2111	2181	2252	2322	70
620	792392	792462	792532	792602	792672	792742	792812	792882	792952	793022	70
1	3092	3162	3231	3301	3371	3441	3511	3581	3651	3721	70
2	3790	3860	3930	4000	4070	4139	4209	4279	4349	4418	70
3	4488	4558	4627	4697	4767	4836	4906	4976	5045	5115	70
4	5185	5254	5324	5393	5463	5532	5602	5672	5741	5811	70
625	5880	5949	6019	6088	6158	6227	6297	6366	6436	6505	69
6	6574	6644	6713	6782	6852	6921	6990	7060	7129	7198	69
7	7268	7337	7406	7475	7545	7614	7683	7752	7821	7890	69
8	7960	8029	8098	8167	8236	8305	8374	8443	8513	8582	69
9	8651	8720	8789	8858	8927	8996	9065	9134	9203	9272	69
630	799341	799409	799478	799547	799616	799685	799754	799823	799892	799961	69
1	800029	800098	800167	800236	800305	800373	800442	800511	800580	800648	69
2	0717	0786	0854	0923	0992	1061	1129	1198	1266	1335	69
3	1404	1472	1541	1609	1678	1747	1815	1884	1952	2021	69
4	2089	2158	2226	2295	2363	2432	2500	2568	2637	2705	68
635	2774	2842	2910	2979	3047	3116	3184	3252	3321	3389	68
6	3457	3525	3594	3662	3730	3798	3867	3935	4003	4071	68
7	4139	4208	4276	4344	4412	4480	4548	4616	4685	4753	68
8	4821	4889	4957	5025	5093	5161	5229	5297	5365	5433	68
9	5501	5569	5637	5705	5773	5841	5908	5976	6044	6112	68
N.	0	1	2	3	4	5	6	7	8	9	D.

N.	0	1	2	3	4	5	6	7	8	9	D.
640	806180	806248	806316	806384	806451	806519	806587	806655	806723	806790	68
1	6858	6926	6994	7061	7129	7197	7264	7332	7400	7467	68
2	7535	7603	7670	7738	7806	7873	7941	8008	8076	8143	68
3	8211	8279	8346	8414	8481	8549	8616	8684	8751	8818	67
4	8886	8953	9021	9088	9156	9223	9290	9358	9425	9492	67
645	9560	9627	9694	9762	9829	9896	9964	810031	810098	810165	67
6	810233	810300	810367	810434	810501	810569	810636	0703	0770	0837	67
7	0904	0971	1039	1106	1173	1240	1307	1374	1441	1508	67
8	1575	1642	1709	1776	1843	1910	1977	2044	2111	2178	67
9	2245	2312	2379	2445	2512	2579	2646	2713	2780	2847	67
650	812913	812980	813047	813114	813181	813247	813314	813381	813448	813514	67
1	3581	3648	3714	3781	3848	3914	3981	4048	4114	4181	67
2	4248	4314	4381	4447	4514	4581	4647	4714	4780	4847	67
3	4913	4980	5046	5113	5179	5246	5312	5378	5445	5511	66
4	5578	5644	5711	5777	5843	5910	5976	6042	6109	6175	66
655	6241	6308	6374	6440	6506	6573	6639	6705	6771	6838	66
6	6904	6970	7036	7102	7169	7235	7301	7367	7433	7499	66
7	7565	7631	7698	7764	7830	7896	7962	8028	8094	8160	66
8	8226	8292	8358	8424	8490	8556	8622	8688	8754	8820	66
9	8885	8951	9017	9083	9149	9215	9281	9346	9412	9478	66
660	819544	819610	819676	819741	819807	819873	819939	820004	820070	820136	66
1	820201	820267	820333	820399	820464	820530	820595	0661	0727	0792	66
2	0858	0924	0989	1055	1120	1186	1251	1317	1382	1448	66
3	1514	1579	1645	1710	1775	1841	1906	1972	2037	2103	65
4	2168	2233	2299	2364	2430	2495	2560	2626	2691	2756	65
665	2822	2887	2952	3018	3083	3148	3213	3279	3344	3409	65
6	3474	3539	3605	3670	3735	3800	3865	3930	3996	4061	65
7	4126	4191	4256	4321	4386	4451	4516	4581	4646	4711	65
8	4776	4841	4906	4971	5036	5101	5166	5231	5296	5361	65
9	5426	5491	5556	5621	5686	5751	5815	5880	5945	6010	65
670	826075	826140	826204	826269	826334	826399	826464	826528	826593	826658	65
1	6723	6787	6852	6917	6981	7046	7111	7175	7240	7305	65
2	7369	7434	7499	7563	7628	7692	7757	7821	7886	7951	65
3	8015	8080	8144	8209	8273	8338	8402	8467	8531	8595	64
4	8660	8724	8789	8853	8918	8982	9046	9111	9175	9239	64
675	9304	9368	9432	9497	9561	9625	9690	9754	9818	9882	64
6	9947	830011	830075	830139	830204	830268	830332	830396	830460	830525	64
7	830589	0653	0717	0781	0845	0909	0973	1037	1102	1166	64
8	1230	1294	1358	1422	1486	1550	1614	1678	1742	1806	64
9	1870	1934	1998	2062	2126	2189	2253	2317	2381	2445	64
680	832509	832573	832637	832700	832764	832828	832892	832956	833020	833083	64
1	3147	3211	3275	3338	3402	3466	3530	3593	3657	3721	64
2	3784	3848	3912	3975	4039	4103	4166	4230	4294	4357	64
3	4421	4484	4548	4611	4675	4739	4802	4866	4929	4993	64
4	5056	5120	5183	5247	5310	5373	5437	5500	5564	5627	63
685	5691	5754	5817	5881	5944	6007	6071	6134	6197	6261	63
6	6324	6387	6451	6514	6577	6641	6704	6767	6830	6894	63
7	6957	7020	7083	7146	7210	7273	7336	7399	7462	7525	63
8	7588	7652	7715	7778	7841	7904	7967	8030	8093	8156	63
9	8219	8282	8345	8408	8471	8534	8597	8660	8723	8786	63
690	838849	838912	838975	839038	839101	839164	839227	839289	839352	839415	63
1	9478	9541	9604	9667	9729	9792	9855	9918	9981	840043	63
2	840106	840169	840232	840294	840357	840420	840482	840545	840608	0671	63
3	0733	0796	0859	0921	0984	1046	1109	1172	1234	1297	63
4	1359	1422	1485	1547	1610	1672	1735	1797	1860	1922	63
695	1985	2047	2110	2172	2235	2297	2360	2422	2484	2547	62
6	2609	2672	2734	2796	2859	2921	2983	3046	3108	3170	62
7	3233	3295	3357	3420	3482	3544	3606	3669	3731	3793	62
8	3855	3918	3980	4042	4104	4166	4229	4291	4353	4415	62
9	4477	4539	4601	4664	4726	4788	4850	4912	4974	5036	62
N.	0	1	2	3	4	5	6	7	8	9	D.

N.	0	1	2	3	4	5	6	7	8	9	D.
700	845098	845160	845222	845284	845346	845408	845470	845532	845594	845656	62
1	5718	5780	5842	5904	5966	6028	6090	6151	6213	6275	62
2	6337	6399	6461	6523	6585	6646	6708	6770	6832	6894	62
3	6955	7017	7079	7141	7202	7264	7326	7338	7449	7511	62
4	7573	7634	7696	7758	7819	7881	7943	8004	8066	8128	62
705	8189	8251	8312	8374	8435	8497	8559	8620	8682	8743	62
6	8805	8866	8928	8989	9051	9112	9174	9235	9297	9358	61
7	9419	9481	9542	9604	9665	9726	9788	9849	9911	9972	61
8	850033	850095	850156	850217	850279	850340	850401	850462	850524	850585	61
9	0646	0707	0769	0830	0891	0952	1014	1075	1136	1197	61
710	851258	851320	851381	851442	851503	851564	851625	851686	851747	851809	61
1	1870	1931	1992	2053	2114	2175	2236	2297	2358	2419	61
2	2480	2541	2602	2663	2724	2785	2846	2907	2968	3029	61
3	3090	3150	3211	3272	3333	3394	3455	3516	3577	3637	61
4	3698	3759	3820	3881	3941	4002	4063	4124	4185	4245	61
715	4306	4367	4428	4488	4549	4610	4670	4731	4792	4852	61
6	4913	4974	5034	5095	5156	5216	5277	5337	5398	5459	61
7	5519	5580	5640	5701	5761	5822	5882	5943	6003	6064	61
8	6124	6185	6245	6306	6366	6427	6487	6548	6608	6668	60
9	6729	6789	6850	6910	6970	7031	7091	7152	7212	7272	60
720	857332	857393	857453	857513	857574	857634	857694	857755	857815	857875	60
1	7935	7995	8056	8116	8176	8236	8297	8357	8417	8477	60
2	8537	8597	8657	8718	8778	8838	8898	8958	9018	9078	60
3	9138	9198	9258	9318	9379	9439	9499	9559	9619	9679	60
4	9739	9799	9859	9918	9978	860038	860098	860158	860218	860278	60
725	860338	860398	860458	860518	860578	0637	0697	0757	0817	0877	60
6	0937	0996	1056	1116	1176	1236	1295	1355	1415	1475	60
7	1534	1594	1654	1714	1773	1833	1893	1952	2012	2072	60
8	2131	2191	2251	2310	2370	2430	2489	2549	2608	2668	60
9	2728	2787	2847	2906	2966	3025	3085	3144	3204	3263	60
730	863323	863382	863442	863501	863561	863620	863680	863739	863799	863858	59
1	3917	3977	4036	4096	4155	4214	4274	4333	4392	4452	59
2	4511	4570	4630	4689	4748	4808	4867	4926	4985	5045	59
3	5104	5163	5222	5282	5341	5400	5459	5519	5578	5637	59
4	5696	5755	5814	5874	5933	5992	6051	6110	6169	6228	59
735	6287	6346	6405	6465	6524	6583	6642	6701	6760	6819	59
6	6878	6937	6996	7055	7114	7173	7232	7291	7350	7409	59
7	7467	7526	7585	7644	7703	7762	7821	7880	7939	7998	59
8	8056	8115	8174	8233	8292	8350	8409	8468	8527	8586	59
9	8644	8703	8762	8821	8879	8938	8997	9056	9114	9173	59
740	869232	869290	869349	869408	869466	869525	869584	869642	839701	869760	59
1	9818	9877	9935	9994	870053	870111	870170	870228	870287	870345	59
2	870404	870462	870521	870579	0638	0696	0755	0813	0872	0930	58
3	0989	1047	1106	1164	1223	1281	1339	1398	1456	1515	58
4	1573	1631	1690	1748	1806	1865	1923	1981	2040	2098	58
745	2156	2215	2273	2331	2389	2448	2506	2564	2622	2681	58
6	2739	2797	2855	2913	2972	3030	3088	3146	3204	3262	58
7	3321	3379	3437	3495	3553	3611	3669	3727	3785	3844	58
8	3902	3960	4018	4076	4134	4192	4250	4308	4366	4424	58
9	4482	4540	4598	4656	4714	4772	4830	4888	4945	5003	58
750	875061	875119	875177	875235	875293	875351	875409	875466	875524	875582	58
1	5640	5698	5756	5813	5871	5929	5987	6045	6102	6160	58
2	6218	6276	6333	6391	6449	6507	6564	6622	6680	6737	58
3	6795	6853	6910	6968	7026	7083	7141	7199	7256	7314	58
4	7371	7429	7487	7544	7602	7659	7717	7774	7832	7889	58
755	7947	8004	8062	8119	8177	8234	8292	8349	8407	8464	57
6	8522	8579	8637	8694	8752	8809	8866	8924	8981	9039	57
7	9096	9153	9211	9268	9325	9383	9440	9497	9555	9612	57
8	9669	9726	9784	9841	9898	9956	880013	880070	880127	880185	57
9	880242	880299	880356	880413	880471	880528	0585	0642	0699	0756	57
N.	0	1	2	3	4	5	6	7	8	9	D.

N.	0	1	2	3	4	5	6	7	8	9	D.
760	880814	880871	880928	880985	881042	881099	881156	881213	881271	881328	57
1	1385	1442	1499	1556	1613	1670	1727	1784	1841	1898	57
2	1955	2012	2069	2126	2183	2240	2297	2354	2411	2468	57
3	2525	2581	2638	2695	2752	2809	2866	2923	2980	3037	57
4	3093	3150	3207	3264	3321	3377	3434	3491	3548	3605	57
765	3661	3718	3775	3832	3888	3945	4002	4059	4115	4172	57
6	4229	4285	4342	4399	4455	4512	4569	4625	4682	4739	57
7	4795	4852	4909	4965	5022	5078	5135	5192	5248	5305	57
8	5361	5418	5474	5531	5587	5644	5700	5757	5813	5870	57
9	5926	5983	6039	6096	6152	6209	6265	6321	6378	6434	56
770	886491	886547	886604	886660	886716	886773	886829	886885	886942	886998	56
1	7054	7111	7167	7223	7280	7336	7392	7449	7505	7561	56
2	7617	7674	7730	7786	7842	7898	7955	8011	8067	8123	56
3	8179	8236	8292	8348	8404	8460	8516	8573	8629	8685	56
4	8741	8797	8853	8909	8965	9021	9077	9134	9190	9246	56
775	9302	9358	9414	9470	9526	9582	9638	9694	9750	9806	56
6	9862	9918	9974	890030	890086	890141	890197	890253	890309	890365	56
7	890421	890477	890533	0589	0645	0700	0756	0812	0868	0924	56
8	0980	1035	1091	1147	1203	1259	1314	1370	1426	1482	56
9	1537	1593	1649	1705	1760	1816	1872	1928	1983	2039	56
780	892095	892150	892206	892262	892317	892373	892429	892484	892540	892595	56
1	2651	2707	2762	2818	2873	2929	2985	3040	3096	3151	56
2	3207	3262	3318	3373	3429	3484	3540	3595	3651	3706	56
3	3762	3817	3873	3928	3984	4039	4094	4150	4205	4261	55
4	4316	4371	4427	4482	4538	4593	4648	4704	4759	4814	55
785	4870	4925	4980	5036	5091	5146	5201	5257	5312	5367	55
6	5423	5478	5533	5588	5644	5699	5754	5809	5864	5920	55
7	5975	6030	6085	6140	6195	6251	6306	6361	6416	6471	55
8	6526	6581	6636	6692	6747	6802	6857	6912	6967	7022	55
9	7077	7132	7187	7242	7297	7352	7407	7462	7517	7572	55
790	897627	897682	897737	897792	897847	897902	897957	898012	898067	898122	55
1	8176	8231	8286	8341	8396	8451	8506	8561	8615	8670	55
2	8725	8780	8835	8890	8944	8999	9054	9109	9164	9218	55
3	9273	9328	9383	9437	9492	9547	9602	9656	9711	9766	55
4	9821	9875	9930	9985	900039	900094	900149	900203	900258	900312	55
795	900367	900422	900476	900531	0586	0640	0695	0749	0804	0859	55
6	0913	0968	1022	1077	1131	1186	1240	1295	1349	1404	55
7	1458	1513	1567	1622	1676	1731	1785	1840	1894	1948	54
8	2003	2057	2112	2166	2221	2275	2329	2384	2438	2492	54
9	2547	2601	2655	2710	2764	2818	2873	2927	2981	3036	54
800	903090	903144	903199	903253	903307	903361	903416	903470	903524	903578	54
1	3633	3687	3741	3795	3849	3904	3958	4012	4066	4120	54
2	4174	4229	4283	4337	4391	4445	4499	4553	4607	4661	54
3	4716	4770	4824	4878	4932	4986	5040	5094	5148	5202	54
4	5256	5310	5364	5418	5472	5526	5580	5634	5688	5742	54
805	5796	5850	5904	5958	6012	6066	6119	6173	6227	6281	54
6	6335	6389	6443	6497	6551	6604	6658	6712	6766	6820	54
7	6874	6927	6981	7035	7089	7143	7196	7250	7304	7358	54
8	7411	7465	7519	7573	7626	7680	7734	7787	7841	7895	54
9	7949	8002	8056	8110	8163	8217	8270	8324	8378	8431	54
810	908485	908539	908592	908646	908699	908753	908807	908860	908914	908967	54
1	9021	9074	9128	9181	9235	9289	9342	9396	9449	9503	54
2	9556	9610	9663	9716	9770	9823	9877	9930	9984	910037	53
3	910091	910144	910197	910251	910304	910358	910411	910464	910518	0571	53
4	0624	0678	0731	0784	0838	0891	0944	0998	1051	1104	53
815	1158	1211	1264	1317	1371	1424	1477	1530	1584	1637	53
6	1690	1743	1797	1850	1903	1956	2009	2063	2116	2169	53
7	2222	2275	2328	2381	2435	2488	2541	2594	2647	2700	53
8	2753	2806	2859	2913	2966	3019	3072	3125	3178	3231	53
9	3284	3337	3390	3443	3496	3549	3602	3655	3708	3761	53
N.	0	1	2	3	4	5	6	7	8	9	D.

N.	0	1	2	3	4	5	6	7	8	9	D.
820	913814	913867	913920	913973	914026	914079	914132	914184	914237	914290	53
1	4343	4396	4449	4502	4555	4608	4660	4713	4766	4819	53
2	4872	4925	4977	5030	5083	5136	5189	5241	5294	5347	53
3	5400	5453	5505	5558	5611	5664	5716	5769	5822	5875	53
4	5927	5980	6033	6085	6138	6191	6243	6296	6349	6401	53
825	6454	6507	6559	6612	6664	6717	6770	6822	6875	6927	53
6	6980	7033	7085	7138	7190	7243	7295	7348	7400	7453	53
7	7506	7558	7611	7663	7716	7768	7820	7873	7925	7978	52
8	8030	8083	8135	8188	8240	8293	8345	8397	8450	8502	52
9	8555	8607	8659	8712	8764	8816	8869	8921	8973	9026	52
830	919078	919130	919183	919235	919287	919340	919392	919444	919496	919549	52
1	9601	9653	9706	9758	9810	9862	9914	9967	920019	920071	52
2	920123	920176	920228	920280	920332	920384	920436	920489	0541	0593	52
3	0645	0697	0749	0801	0853	0906	0958	1010	1062	1114	52
4	1166	1218	1270	1322	1374	1426	1478	1530	1582	1634	52
835	1686	1738	1790	1842	1894	1946	1998	2050	2102	2154	52
6	2206	2258	2310	2362	2414	2466	2518	2570	2622	2674	52
7	2725	2777	2829	2881	2933	2985	3037	3089	3140	3192	52
8	3244	3296	3348	3399	3451	3503	3555	3607	3658	3710	52
9	3762	3814	3865	3917	3969	4021	4072	4124	4176	4228	52
840	924279	924331	924383	924434	924486	924538	924589	924641	924693	924744	52
1	4796	4848	4899	4951	5003	5054	5106	5157	5209	5261	52
2	5312	5364	5415	5467	5518	5570	5621	5673	5725	5776	52
3	5828	5879	5931	5982	6034	6085	6137	6188	6240	6291	51
4	6342	6394	6445	6497	6548	6600	6651	6702	6754	6805	51
845	6857	6908	6959	7011	7062	7114	7165	7216	7268	7319	51
6	7370	7422	7473	7524	7576	7627	7678	7730	7781	7832	51
7	7883	7935	7986	8037	8088	8140	8191	8242	8293	8345	51
8	8396	8447	8498	8549	8601	8652	8703	8754	8805	8857	51
9	8908	8959	9010	9061	9112	9163	9215	9266	9317	9368	51
850	929419	929470	929521	929572	929623	929674	929725	929776	929827	929879	51
1	9930	9981	930032	930083	930134	930185	930236	930287	930338	930389	51
2	930440	930491	0542	0592	0643	0694	0745	0796	0847	0898	51
3	0949	1000	1051	1102	1153	1204	1254	1305	1356	1407	51
4	1458	1509	1560	1610	1661	1712	1763	1814	1865	1915	51
855	1966	2017	2068	2118	2169	2220	2271	2322	2372	2423	51
6	2474	2524	2575	2626	2677	2727	2778	2829	2879	2930	51
7	2981	3031	3082	3133	3183	3234	3285	3335	3386	3437	51
8	3487	3538	3589	3639	3690	3740	3791	3841	3892	3943	51
9	3993	4044	4094	4145	4195	4246	4296	4347	4397	4448	51
860	934498	934549	934599	934650	934700	934751	934801	934852	934902	934953	50
1	5003	5054	5104	5154	5205	5255	5306	5356	5406	5457	50
2	5507	5558	5608	5658	5709	5759	5809	5860	5910	5960	50
3	6011	6061	6111	6162	6212	6262	6313	6363	6413	6463	50
4	6514	6564	6614	6665	6715	6765	6815	6865	6916	6966	50
865	7016	7066	7117	7167	7217	7267	7317	7367	7418	7468	50
6	7518	7568	7618	7668	7718	7769	7819	7869	7919	7969	50
7	8019	8069	8119	8169	8219	8269	8320	8370	8420	8470	50
8	8520	8570	8620	8670	8720	8770	8820	8870	8920	8970	50
9	9020	9070	9120	9170	9220	9270	9320	9369	9419	9469	50
870	939519	939569	939619	939669	939719	939769	939819	939869	939918	939968	50
1	940018	940068	940118	940168	940218	940267	940317	940367	940417	940467	50
2	0516	0566	0616	0666	0716	0765	0815	0865	0915	0964	50
3	1014	1064	1114	1163	1213	1263	1313	1362	1412	1462	50
4	1511	1561	1611	1660	1710	1760	1809	1859	1909	1958	50
875	2008	2058	2107	2157	2207	2256	2306	2355	2405	2455	50
6	2504	2554	2603	2653	2702	2752	2801	2851	2901	2950	50
7	3000	3049	3099	3148	3198	3247	3297	3346	3396	3445	49
8	3495	3544	3593	3643	3692	3742	3791	3841	3890	3939	49
9	3989	4038	4088	4137	4186	4236	4285	4335	4384	4433	49
N.	0	1	2	3	4	5	6	7	8	9	D.

N.	0	1	2	3	4	5	6	7	8	9	D.
880	944483	944532	944581	944631	944680	944729	944779	944828	944877	944927	49
1	4976	5025	5074	5124	5173	5222	5272	5321	5370	5419	49
2	5469	5518	5567	5616	5665	5715	5764	5813	5862	5912	49
3	5961	6010	6059	6108	6157	6207	6256	6305	6354	6403	49
4	6452	6501	6551	6600	6649	6698	6747	6796	6845	6894	49
885	6943	6992	7041	7090	7140	7189	7238	7287	7336	7385	49
6	7434	7483	7532	7581	7630	7679	7728	7777	7826	7875	49
7	7924	7973	8022	8070	8119	8168	8217	8266	8315	8364	49
8	8413	8462	8511	8560	8609	8657	8706	8755	8804	8853	49
9	8902	8951	8999	9048	9097	9146	9195	9244	9292	9341	49
890	949390	949439	949488	949536	949585	949634	949683	949731	949780	949829	49
1	9878	9926	9975	950024	950073	950121	950170	950219	950267	950316	49
2	950365	950414	950462	0511	0560	0608	0657	0706	0754	0803	49
3	0851	0900	0949	0997	1046	1095	1143	1192	1240	1289	49
4	1338	1386	1435	1483	1532	1580	1629	1677	1726	1775	49
895	1823	1872	1920	1969	2017	2066	2114	2163	2211	2260	48
6	2308	2356	2405	2453	2502	2550	2599	2647	2696	2744	48
7	2792	2841	2889	2938	2986	3034	3083	3131	3180	3228	48
8	3276	3325	3373	3421	3470	3518	3566	3615	3663	3711	48
9	3760	3808	3856	3905	3953	4001	4049	4098	4146	4194	48
900	954243	954291	954339	954387	954435	954484	954532	954580	954628	954677	48
1	4725	4773	4821	4869	4918	4966	5014	5062	5110	5158	48
2	5207	5255	5303	5351	5399	5447	5495	5543	5592	5640	48
3	5688	5736	5784	5832	5880	5928	5976	6024	6072	6120	48
4	6168	6216	6265	6313	6361	6409	6457	6505	6553	6601	48
905	6649	6697	6745	6793	6840	6888	6936	6984	7032	7080	48
6	7128	7176	7224	7272	7320	7368	7416	7464	7512	7559	48
7	7607	7655	7703	7751	7799	7847	7894	7942	7990	8038	48
8	8086	8134	8181	8229	8277	8325	8373	8421	8468	8516	48
9	8564	8612	8659	8707	8755	8803	8850	8898	8946	8994	48
910	959041	959089	959137	959185	959232	959280	959328	959375	959423	959471	48
1	9518	9566	9614	9661	9709	9757	9804	9852	9900	9947	48
2	9995	960042	960090	960138	960185	960233	960280	960328	960376	960423	48
3	960471	0518	0566	0613	0661	0709	0756	0804	0851	0899	48
4	0946	0994	1041	1089	1136	1184	1231	1279	1326	1374	48
915	1421	1469	1516	1563	1611	1658	1706	1753	1801	1848	47
6	1895	1943	1990	2038	2085	2132	2180	2227	2275	2322	47
7	2369	2417	2464	2511	2559	2606	2653	2701	2748	2795	47
8	2843	2890	2937	2985	3032	3079	3126	3174	3221	3268	47
9	3316	3363	3410	3457	3504	3552	3599	3646	3693	3741	47
920	963788	963835	963882	963929	963977	964024	964071	964118	964165	964212	47
1	4260	4307	4354	4401	4448	4495	4542	4590	4637	4684	47
2	4731	4778	4825	4872	4919	4966	5013	5061	5108	5155	47
3	5202	5249	5296	5343	5390	5437	5484	5531	5578	5625	47
4	5672	5719	5766	5813	5860	5907	5954	6001	6048	6095	47
925	6142	6189	6236	6283	6329	6376	6423	6470	6517	6564	47
6	6611	6658	6705	6752	6799	6845	6892	6939	6986	7033	47
7	7080	7127	7173	7220	7267	7314	7361	7408	7454	7501	47
8	7548	7595	7642	7688	7735	7782	7829	7875	7922	7969	47
9	8016	8062	8109	8156	8203	8249	8296	8343	8390	8436	47
930	968483	968530	968576	968623	968670	968716	968763	968810	968856	968903	47
1	8950	8996	9043	9090	9136	9183	9229	9276	9323	9369	47
2	9416	9463	9509	9556	9602	9649	9695	9742	9789	9835	47
3	9882	9928	9975	970021	970068	970114	970161	970207	970254	970300	47
4	970347	970393	970440	0486	0533	0579	0626	0672	0719	0765	46
935	0812	0858	0904	0951	0997	1044	1090	1137	1183	1229	46
6	1276	1322	1369	1415	1461	1508	1554	1601	1647	1693	46
7	1740	1786	1832	1879	1925	1971	2018	2064	2110	2157	46
8	2203	2249	2295	2342	2388	2434	2481	2527	2573	2619	46
9	2666	2712	2758	2804	2851	2897	2943	2989	3035	3082	46
N.	0	1	2	3	4	5	6	7	8	9	D.

N.	0	1	2	3	4	5	6	7	8	9	D.
940	973128	973174	973220	973266	973313	973359	973405	973451	973497	973543	46
1	3590	3636	3682	3728	3774	3820	3866	3913	3959	4005	46
2	4051	4097	4143	4189	4235	4281	4327	4374	4420	4466	46
3	4512	4558	4604	4650	4696	4742	4788	4834	4880	4926	46
4	4972	5018	5064	5110	5156	5202	5248	5294	5340	5386	46
945	5432	5478	5524	5570	5616	5662	5707	5753	5799	5845	46
6	5891	5937	5983	6029	6075	6121	6167	6212	6258	6304	46
7	6350	6396	6442	6488	6533	6579	6625	6671	6717	6763	46
8	6808	6854	6900	6946	6992	7037	7083	7129	7175	7220	46
9	7266	7312	7358	7403	7449	7495	7541	7586	7632	7678	46
950	977724	977769	977815	977861	977906	977952	977998	978043	978089	978135	46
1	8181	8226	8272	8317	8363	8409	8454	8500	8546	8591	46
2	8637	8683	8728	8774	8819	8865	8911	8956	9002	9047	46
3	9093	9138	9184	9230	9275	9321	9366	9412	9457	9503	46
4	9548	9594	9639	9685	9730	9776	9821	9867	9912	9958	46
955	980003	980049	980094	980140	980185	980231	980276	980322	980367	980412	45
6	0458	0503	0549	0594	0640	0685	0730	0776	0821	0867	45
7	0912	0957	1003	1048	1093	1139	1184	1229	1275	1320	45
8	1366	1411	1456	1501	1547	1592	1637	1683	1728	1773	45
9	1819	1864	1909	1954	2000	2045	2090	2135	2181	2226	45
960	982271	982316	982362	982407	982452	982497	982543	982588	982633	982678	45
1	2723	2769	2814	2859	2904	2949	2994	3040	3085	3130	45
2	3175	3220	3265	3310	3356	3401	3446	3491	3536	3581	45
3	3626	3671	3716	3762	3807	3852	3897	3942	3987	4032	45
4	4077	4122	4167	4212	4257	4302	4347	4392	4437	4482	45
965	4527	4572	4617	4662	4707	4752	4797	4842	4887	4932	45
6	4977	5022	5067	5112	5157	5202	5247	5292	5337	5382	45
7	5426	5471	5516	5561	5606	5651	5696	5741	5786	5830	45
8	5875	5920	5965	6010	6055	6100	6144	6189	6234	6279	45
9	6324	6369	6413	6458	6503	6548	6593	6637	6682	6727	45
970	986772	986817	986861	986906	986951	986996	987040	987085	987130	987175	45
1	7219	7264	7309	7353	7398	7443	7488	7532	7577	7622	45
2	7666	7711	7756	7800	7845	7890	7934	7979	8024	8068	45
3	8113	8157	8202	8247	8291	8336	8381	8425	8470	8514	45
4	8559	8604	8648	8693	8737	8782	8826	8871	8916	8960	45
975	9005	9049	9094	9138	9183	9227	9272	9316	9361	9405	45
6	9450	9494	9539	9583	9628	9672	9717	9761	9806	9850	44
7	9395	9939	9983	990028	990072	990117	990161	990206	990250	990294	44
8	990339	990383	990428	0472	0516	0561	0605	0650	0694	0738	44
9	0783	0827	0871	0916	0960	1004	1049	1093	1137	1182	44
980	991226	991270	991315	991359	991403	991448	991492	991536	991580	991625	44
1	1669	1713	1758	1802	1846	1890	1935	1979	2023	2067	44
2	2111	2156	2200	2244	2288	2333	2377	2421	2465	2509	44
3	2554	2598	2642	2686	2730	2774	2819	2863	2907	2951	44
4	2995	3039	3083	3127	3172	3216	3260	3304	3348	3392	44
985	3436	3480	3524	3568	3613	3657	3701	3745	3789	3833	44
6	3877	3921	3965	4009	4053	4097	4141	4185	4229	4273	44
7	4317	4361	4405	4449	4493	4537	4581	4625	4669	4713	44
8	4757	4801	4845	4889	4933	4977	5021	5065	5108	5152	44
9	5196	5240	5284	5328	5372	5416	5460	5504	5547	5591	44
990	995635	995679	995723	995767	995811	995854	995898	995942	995986	996030	44
1	6074	6117	6161	6205	6249	6293	6337	6380	6424	6468	44
2	6512	6555	6599	6643	6687	6731	6774	6818	6862	6906	44
3	6949	6993	7037	7080	7124	7168	7212	7255	7299	7343	44
4	7386	7430	7474	7517	7561	7605	7648	7692	7736	7779	44
995	7823	7867	7910	7954	7998	8041	8085	8129	8172	8216	44
6	8259	8303	8347	8390	8434	8477	8521	8564	8608	8652	44
7	8695	8739	8782	8826	8869	8913	8956	9000	9043	9087	44
8	9131	9174	9218	9261	9305	9348	9392	9435	9479	9522	44
9	9565	9609	9652	9696	9739	9783	9826	9870	9913	9957	43
N.	0	1	2	3	4	5	6	7	8	9	D.

APPENDIX Q

Glossary of Symbols and Formulae

For the convenience of the reader, the more important symbols and formulae are listed below. The arrangement is alphabetical, and where Greek characters are shown, they are arranged according to the way they are pronounced in English. Occasionally a given symbol has more than one meaning, but the meaning intended is indicated by the context.

In formulae having to do with multiple and partial correlation, various combinations of subscripts are possible with a, b, β, d, R, r, σ_C, σ_S, Σx_C^2, Σx_S^2 The general practice in this glossary is to give the formula for a specific combination, such as $r_{12\,34}$ The reader can easily supply the corresponding formulae for $r_{13.24}$, $r_{14\,23}$, etc.

Greek Alphabet

Greek		English	Greek		English	Greek		English
A	α	Alpha	I	ι	Iota	P	ρ	Rho
B	β	Beta	K	κ	Kappa	Σ	σ	Sigma
Γ	γ	Gamma	Λ	λ	Lambda	T	τ	Tau
Δ	δ	Delta	M	μ	Mu	Υ	υ	Upsilon
E	ϵ	Epsilon	N	ν	Nu	Φ	ϕ	Phi
Z	ς	Zeta	Ξ	ξ	Xi	X	χ	Cni
H	η	Eta	O	o	Omicron	Ψ	$\dot\psi$	Psi
Θ	θ	Theta	Π	π	Pi	Ω	ω	Omega

$A = \dfrac{\Sigma Y}{N}$: a constant in an orthogonal polynomial equation.

$A = \dfrac{2}{T}\Sigma\left[Y \sin\left(\dfrac{360}{T}X\right)^{\circ}\right]$: a constant in a sine cosine curve.

$a = \Sigma_2 Y - \Sigma_1 Y \dfrac{b-1}{(b^n-1)^2}$: a constant in a modified exponential curve

The value of $k - Y_C$ when $X = 0$.

a: a constant in a Gompertz curve equation.

$\log a = (\Sigma_2 \log Y - \Sigma_1 \log Y)\dfrac{b-1}{(b^n-1)^2}$.

a: see a, b, c. d, $e \cdots$ (constants in a polynomial equation).

$a = \log_e \dfrac{k - y_o}{y_o}$: a constant in a Pearl-Reed (logistic) curve.

a: the value of the dependent variable (Y_C) in a polynomial equation when the independent variable (X) is zero. For a straight line equation, $a = \overline{Y} - b\overline{X}$; or, when $\overline{X} = 0$, $a = \dfrac{\Sigma Y}{N}$.

See also a, b, c, d, $e \cdots$: constants in a polynomial equation.

A, B: constants in a sine cosine curve. See also A and B.

A, B, C, $D \cdots$: constants of an orthogonal polynomial equation. For general expression for these constants, see Orthogonal polynomial equation.

a, b, c, $\bar{a}$, $e \cdots$: constants in a polynomial equation.

Normal equations

I. $\Sigma Y = Na + b\Sigma X + c\Sigma X^2 + d\Sigma X^3 + e\Sigma X^4$.
II. $\Sigma XY = a\Sigma X + b\Sigma X^2 + c\Sigma X^3 + d\Sigma X^4 + e\Sigma X^5$.
III. $\Sigma X^2 Y = a\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4 + d\Sigma X^5 + e\Sigma X^6$.
IV. $\Sigma X^3 Y = a\Sigma X^3 + b\Sigma X^4 + c\Sigma X^5 + d\Sigma X^6 + e\Sigma X^7$.
V. $\Sigma X^4 Y = a\Sigma X^4 + b\Sigma X^5 + c\Sigma X^6 + d\Sigma X^7 + e\Sigma X^8$.

When X values are taken as deviations from their mean:

I. $\Sigma Y = Na + c\Sigma X^2 + e\Sigma X^4$.
II. $\Sigma XY = b\Sigma X^2 + d\Sigma X^4$.
III. $\Sigma X^2 Y = a\Sigma X^2 + c\Sigma X^4 + e\Sigma X^6$.
IV. $\Sigma X^3 Y = b\Sigma X^4 + d\Sigma X^6$.
V. $\Sigma X^4 Y = a\Sigma X^4 + c\Sigma X^6 + e\Sigma X^8$.

a, b, k: constants in a modified exponential equation, or in Gompertz or logistic equation employing modified exponential form. See also a, b, and k.

$a_{1\ 234} = \overline{X}_1 - b_{12\ 34}\overline{X}_2 - b_{13\ 24}\overline{X}_3 - b_{14\ 23}\overline{X}_4$: the computed value of the dependent variable ($X_{C1\ 234}$) when the independent variables (X_2, X_3, X_4) are zero.

$a_{1.234}$, $b_{12\ 34}$, $b_{13\ 23}$, $b_{14.23}$: constants in multiple estimating equation.

Normal equations

I. $\Sigma X_1 = Na_{1.234} + b_{12.34}\Sigma X_2 + b_{13\ 24}\Sigma X_3 + b_{14.23}\Sigma X_4$.
II. $\Sigma X_1 X_2 = a_{1\ 234}\Sigma X_2 + b_{12.34}\Sigma X_2^2 + b_{13\ 24}\Sigma X_2 X_3 + b_{14\ 23}\Sigma X_2 X_4$.
III. $\Sigma X_1 X_3 = a_{1\ 234}\Sigma X_3 + b_{12\ 34}\Sigma X_2 X_3 + b_{13.24}\Sigma X_3^2 + b_{14\ 23}\Sigma X_3 X_4$.
IV. $\Sigma X_1 X_4 = a_{1\ 234}\Sigma X_4 + b_{12.34}\Sigma X_2 X_4 + b_{13.24}\Sigma X_3 X_4 + b_{14\ 23}\Sigma X_4^2$.

or

II. $\Sigma x_1 x_2 = b_{12.34}\Sigma x_2^2 + b_{13.24}\Sigma x_2 x_3 + b_{14.23}\Sigma x_2 x_4$.
III. $\Sigma x_1 x_3 = b_{12.34}\Sigma x_2 x_3 + b_{13.24}\Sigma x_3^2 + b_{14.23}\Sigma x_3 x_4$.
IV. $\Sigma x_1 x_4 = b_{12.34}\Sigma x_2 x_4 + b_{13.24}\Sigma x_3 x_4 + b_{14.23}\Sigma x_4^2$

$AD = \dfrac{\Sigma |x|}{N}$, where $\Sigma|x|$ means sum of deviations from mean, signs neg-

lected: average deviation or mean deviation, a measure of absolute dispersion.

Aggregative index number formulae (subscripts to P's and Q's in terms below are for purposes of identification of recurring formulae):

Simple aggregative:

$$P = \frac{\Sigma p_n}{\Sigma p_o}. \qquad Q = \frac{\Sigma q_n}{\Sigma q_o}.$$

Weighted aggregative (general form):

$$P = \frac{\Sigma p_n q}{\Sigma p_o q}. \qquad Q = \frac{\Sigma q_n p}{\Sigma q_o p}$$

Base year weights:.. ..   .   . **1**

$$P_1 = \frac{\Sigma p_n q_o}{\Sigma p_o q_o}. \qquad Q_1 = \frac{\Sigma q_n p_o}{\Sigma q_o p_o}.$$

Given year weights: **2**

$$P_2 = \frac{\Sigma p_n q_n}{\Sigma p_o q_n}. \qquad Q_2 = \frac{\Sigma q_n p_n}{\Sigma q_o p_n}.$$

Marshall-Edgeworth:

$$P = \frac{\Sigma p_n (q_o + q_n)}{\Sigma p_o (q_o + q_n)}, \text{ or } \frac{\Sigma p_n q_{o,n}}{\Sigma p_o q_{o,n}}.$$

Average year weights:

$$P = \frac{\Sigma p_n q_{o-n}}{\Sigma p_o q_{o-n}}.$$

Keynes' common factor:

$$P = \frac{\Sigma p_n q_c}{\Sigma p_o q_c}.$$

"Ideal" index number formula **3**

$$P_3 = \sqrt{P_1 \times P_2} \qquad\qquad Q_3 = \sqrt{Q_1 \times Q_2}$$

$$= \sqrt{\frac{\Sigma p_n q_o}{\Sigma p_o q_o} \times \frac{\Sigma p_n q_n}{\Sigma p_o q_n}}. \qquad = \sqrt{\frac{\Sigma q_n p_o}{\Sigma q_o p_o} \times \frac{\Sigma q_n p_n}{\Sigma q_o p_n}}.$$

Alienation coefficient: see $k = \dfrac{\sigma_{y_s}}{\sigma_y}.$

α: a type of measure describing a frequency distribution. The α's may be computed from the π's, as indicated below; or, in a similar manner from the μ's.

$$\alpha_1 = \frac{\pi_1}{\sigma} = \frac{\pi_1}{\sqrt{\pi_2}} = 0$$

$$\alpha_2 = \frac{\pi_2}{\sigma^2} = \frac{\pi_2}{\sqrt{\pi_2^2}} = 1.$$

$$\alpha_3 = \frac{\pi_3}{\sigma^3} = \frac{\pi_3}{\sqrt{\pi_2^3}} = \sqrt{\beta_1}$$: a measure of relative skewness. α_3 is zero for a normal curve.

$$\alpha_4 = \frac{\pi_4}{\sigma^4} = \frac{\pi_4}{\sqrt{\pi_2^4}} = \frac{\pi_4}{\pi_2^2} = \beta_2$$: a measure of relative kurtosis. α_4 is 3 for a normal curve.

Arithmetic mean: see $\overline{X}$.

Average: see Measure of central tendency.

Average deviation: see AD.

Average of relatives index number formulae:

Simple arithmetic mean:

$$P = \frac{\Sigma\left(\dfrac{p_n}{p_o}\right)}{N}. \qquad Q = \frac{\Sigma\left(\dfrac{q_n}{q_o}\right)}{N}.$$

Simple harmonic mean:

$$P = \frac{N}{\Sigma\left(\dfrac{p_o}{p_n}\right)}. \qquad Q = \frac{N}{\Sigma\left(\dfrac{q_o}{q_n}\right)}.$$

Simple geometric mean:

$$P = \sqrt[N]{\frac{p_n'}{p_o'} \times \frac{p_n''}{p_o''} \times \frac{p_n'''}{p_o'''} \times \cdots}. \qquad Q = \sqrt[N]{\frac{q_n'}{q_o'} \times \frac{q_n''}{q_o''} \times \frac{q_n'''}{q_o'''} \times \cdots}.$$

Weighted arithmetic mean (general form):

$$P = \frac{\Sigma\left(v\,\dfrac{p_n}{p_o}\right)}{\Sigma v}, \qquad Q = \frac{\Sigma\left(v\,\dfrac{q_n}{q_o}\right)}{\Sigma v},$$

where v is $p \times q$.

Arithmetic mean, base year value weights. 4

$$P_4 = \frac{\Sigma\left[(p_o q_o)\,\dfrac{p_n}{p_o}\right]}{\Sigma p_o q_o}. \qquad Q_4 = \frac{\Sigma\left[(q_o p_o)\,\dfrac{q_n}{q_o}\right]}{\Sigma q_o p_o}.$$

(Same as 1.)

Arithmetic mean, mixed weights. 5

$$P_5 = \frac{\Sigma\left[(p_o q_n)\,\dfrac{p_n}{p_o}\right]}{\Sigma p_o q_n}. \qquad Q_5 = \frac{\Sigma\left[(q_o p_n)\,\dfrac{q_n}{q_o}\right]}{\Sigma q_o p_n}.$$

(Same as 2.)

Harmonic mean, given year value weights 6

$$P_6 = \frac{\Sigma p_n q_n}{\Sigma \left[(p_n q_n) \dfrac{p_o}{p_n} \right]}. \qquad Q_6 = \frac{\Sigma q_n p_n}{\Sigma \left[(q_n p_n) \dfrac{q_o}{q_n} \right]}.$$

(Same as 2.)

Harmonic mean, mixed weights 7

$$P_7 = \frac{\Sigma p_n q_o}{\Sigma \left[(p_n q_o) \dfrac{p_o}{p_n} \right]}. \qquad Q_7 = \frac{\Sigma q_n p_o}{\Sigma \left[(q_n p_o) \dfrac{q_o}{q_n} \right]}.$$

(Same as 1.)

Weighted geometric means:

$$P = \sqrt[\Sigma v]{\left(\frac{p_n'}{p_o'}\right)^{v'} \times \left(\frac{p_n''}{p_o''}\right)^{v''} \times \cdots} \qquad Q = \sqrt[\Sigma v]{\left(\frac{q_n'}{q_o'}\right)^{v'} \times \left(\frac{q_n''}{q_o''}\right)^{v''} \times \cdot \cdot}$$

where $v = p_o q$. where $v = q_o p$.

"Ideal" index number formulae 8

$$P_8 = \sqrt{P_4 \times P_6} \qquad\qquad Q_8 = \sqrt{Q_4 \times Q_6}$$

$$\text{or } \sqrt{P_5 \times P_7}. \qquad\qquad\qquad \text{or } \sqrt{Q_5 \times Q_7}.$$

(Same as 3.)

$B = \dfrac{12}{N(N^2 - 1)} \Sigma X_1 Y$: a constant in an orthogonal polynomial equation
See also X_1.

$B = \dfrac{2}{T} \Sigma \left[Y \cos \left(\dfrac{360}{T} X \right)^{\circ} \right]$: a constant in a sine cosine curve.

$b = \sqrt[n]{\dfrac{\Sigma_3 Y - \Sigma_2 Y}{\Sigma_2 Y - \Sigma_1 Y}}$: a constant in a modified exponential equation, the ratio between successive first differences.

$b = \sqrt[n]{\dfrac{\Sigma_3 \log Y - \Sigma_2 \log Y}{\Sigma_2 \log Y - \Sigma_1 \log Y}}$: a constant in a Gompertz curve equation.

$b = \dfrac{1}{n} \log_e \dfrac{y_o (k - y_1)}{y_1 (k - y_o)}$: a constant in a Pearl-Reed (logistic) curve.

b: the slope of the line in a polynomial equation. See $a, b, c, d, e \cdot \cdot$ (constants in a polynomial equation). For a straight line equation,

$$b = \frac{\Sigma XY - \overline{X}\Sigma Y}{\Sigma X^2 - \overline{X}\Sigma X}; \text{ or, when } \overline{X} = 0, b = \frac{\Sigma XY}{\Sigma X^2}.$$

$b = \dfrac{\sigma_P}{\sigma}$ (where σ_P refers to the population): $b\sigma$ gives a fiduciary limit for σ_P

$b_{xy} = \dfrac{\Sigma xy}{\Sigma y^2} = r \dfrac{\sigma_x}{\sigma_y}$: slope of estimating equation $X_C = a' + b'Y$.

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = r\frac{\sigma_y}{\sigma_x}:$$ slope of estimating equation $Y_C = a + bX$.

$$b_{12.34} = r_{12.34}\frac{\sigma_{S1\ 234}}{\sigma_{S2\ 134}}:$$ a coefficient of partial estimation. See also $a_{1\ 234},$

$b_{12\ 34}, b_{13\ 24}, b_{14\ 23}$ (constants in a multiple estimating equation).

β: a criterion of frequency curve type. The β's may be computed from the π's, as indicated below; or, in a similar manner, from the μ's.

$$\cdot\beta_1 = \alpha_3^2 = \frac{\pi_3^2}{\pi_2^3}:$$ a measure of relative skewness. Value of β_1 is 0 for a normal curve.

$$\sqrt{\beta_1} = \alpha_3 = \frac{\pi_3}{\sigma^3} = \frac{\pi_3}{\sqrt{\pi_2^3}}:$$ a measure of relative skewness.

$$\beta_2 = \alpha_4 = \frac{\pi_4}{\sigma^4} = \frac{\pi_4}{\sqrt{\pi_2^4}} = \frac{\pi_4}{\pi_2^2}:$$ a measure of relative kurtosis. Value of β_2 is 3 for a normal curve.

$$\beta_{12\ 34} = b_{12.34}\frac{\sigma_2}{\sigma_1}:$$ a beta coefficient. A measure of the individual importance of one of three independent variables.

Binomial: $(p + q)^m$ for fitting to discrete data, symmetrical if $p = q$. otherwise skewed.

Binomial theorem:

$$(a+b)^m = a^m + ma^{m-1}b + \frac{m(m-1)}{1\ 2}a^{m-2}b^2 + \frac{m(m-1)(m-2)}{1\cdot2\cdot3}a^{m-3}b^3 + \cdots + b^m.$$

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}}:$$ coefficient of mean square contingency, a measure of correlation of qualitatively classified data.

C: computed value. Used only as a subscript in this sense.

C: cyclical movement.

$$C = \frac{180}{N(N^2 - 1)(N^2 - 4)}\Sigma X_2 Y:$$ a constant in an orthogonal polynomial equation.

$$C_2 = \frac{\Sigma s(2N - \Sigma |s|)}{N^2},$$ where s refers to the smaller of each pair of items,

and when each series is expressed as deviations from its mean in terms of its average deviation: first moment correlation coefficient.

c: a correction factor. Formula depends on what is being corrected.

c: change in the slope of a polynomial equation. See $a, b, c, d, e \cdots$ (constants in a polynomial equation).

Camp-Meidell inequality: $P < \dfrac{1}{2.25\left(\dfrac{x}{\sigma}\right)^2},$ where P is the proportion of

items beyond any distance x on both sides of the mean.   Applies to uni-modal distribution if mode is within 1σ of the mean.

χ: same as Sk_β. χ is not used in this text with this meaning.

$\chi: \sqrt{\chi^2}$.

χ^2: a measure of the discrepancy between observed values and theoretical values.

$$\chi^2 = \Sigma \frac{(f - fc)^2}{fc}.$$

$$\chi^2 = \frac{\left(a - \dfrac{p}{q} b\right)^2}{\dfrac{p}{q} N}, \text{ where } a = \text{ number of occurrences of first category:}$$

b = number of occurrences of second category; p = probability of obtaining an occurrence of first category; q = probability of obtaining an occurrence of second category.

$$\chi^2 = \frac{N\sigma^2}{\sigma_{\hat{p}}^2}, \text{ where } \sigma_P \text{ refers to the variance in the population.}$$

Compound interest curve: see Exponential equation.

Correlation, coefficient of: see r.

Correlation, first moment coefficient: see C_2.

Correlation, index of: see ρ.

Correlation, multiple: see $R_{1.234}$.

Correlation, part: see $_{12}r_{34}$.

Correlation, partial: see $r_{12.34}$.

Correlation ratio: see η.

Cyclical-irregular movements: $C \times I$.

D: a decile.   There are nine deciles, $D_1 \cdots D_9$.

D: a difference between paired values.

$$D = \frac{2,800}{N(N^2 - 1)(N^2 - 4)(N^2 - 9)} \Sigma X_3 Y: \text{ a constant in an orthogonal}$$
polynomial equation.

d: a constant in a polynomial equation.   See $a, b, c, d, e \cdots$ (constants in a polynomial equation).

$d = X - \overline{X}_a$: a deviation from an assumed mean.

$$d' = \frac{X - \overline{X}_d}{i}: \text{ a deviation from an assumed mean in units of class intervals}$$

$$d_{12\ 34}^2 = \frac{b_{12\ 34}\Sigma x_1 x_2}{\Sigma x_1^2}: \text{ a coefficient of separate determination.}$$

Δ: a finite difference.

Δ^1 is a first difference; Δ^2 is a second difference.

Δ_1, as used in computation of the mode, is the arithmetic difference between the frequency of the modal class and the frequency of the preceding class; Δ_2 is the arithmetic difference between the frequency of the modal class and the frequency of the following class.

Deseasonalized data: $T \times C \times I$.

Determination, coefficient of multiple: see R_1^2 234.

Determination, coefficient of partial: see $r_{12\ 34}^2$.

Determination, coefficient of separate: see $d_{12\ 34}^2$.

Determination, coefficient of simple: see r^2.

Determination, index of, see ρ^2.

Determination, ratio of, see η^2.

Dispersion:

For absolute dispersion see σ; $\pi_2 = \sigma^2$; AD; Q.

For relative dispersion see V.

e: a constant in a polynomial equation. See a, b, c, d, $e \cdot \cdot \cdot$ (constants in a polynomial equation).

$e = 2.71828$: the base of the Naperian, or natural, logarithmic system.

The limit of $(1 + \dfrac{1}{n})^n$. $\left(\log_e X = \dfrac{\log_{10} X}{.43429} = 2.30259 \log_{10} X.\right)$

η (correlation ratio): see η^2.

$$\eta^2 = \frac{\overset{m}{\underset{1}{\Sigma}}\left[N_K(\bar{Y}_K - \bar{Y})^2\right]}{\Sigma(Y - \bar{Y})^2} = \frac{\overset{m}{\underset{1}{\Sigma}}\left(\bar{Y}_K \overset{N_K}{\underset{1}{\Sigma}} Y\right) - \bar{Y}\Sigma Y}{\Sigma Y^2 - \bar{Y}\Sigma Y} : \text{ratio of determination.}$$

A measure of relationship when data are grouped along X-axis and line of means is taken as estimating line.

$$\bar{\eta}^2 = \frac{\eta^2 (N - 1) - (m - 1)}{N - m} : \text{estimated population value of } \eta^2.$$

Explained sum of squares:

For simple correlation (linear and non-linear):

$$\Sigma Y_{\bar{c}}^2 = a\Sigma Y + b\Sigma XY + c\Sigma X^2 Y + d\Sigma X^3 Y + \cdot \cdot \cdot.$$

For simple correlation grouped data in units of class intervals (deviations are from assumed means):

$$\Sigma f_y (d'_{y_c})^2 = a\Sigma f_y\, d'_y + b\Sigma f d'_x\, d'_y + \cdot \cdot \cdot.$$

For multiple correlation:

$$\Sigma X_{C1.234}^2 = a_{1.234}\Sigma X_1 + b_{12.34}\Sigma X_1 X_2 + b_{13.24}\Sigma X_1 X_3 + b_{14.23}\Sigma X_1 X_4$$

Explained variance: see $\sigma_{y_c}^2$.

Explained variation (see also Explained sum of squares):

For simple correlation:

$$\Sigma y_{\bar{c}}^2 = \Sigma (Y_C - \bar{Y})^2 = \Sigma Y_{\bar{c}}^2 - \bar{Y}\Sigma Y.$$

For simple correlation when data are in deviation form:

$$\Sigma y_C^2 = b\Sigma xy + c\Sigma x^2 y + d\Sigma x^3 y + \cdots .$$

For simple correlation, grouped data (see also Variation between columns):

$$\Sigma f y_C^2 = i^2 \left[\Sigma f_y (d'_{y_c})^2 - \frac{(\Sigma f_y d'_y)^2}{N} \right].$$

For simple correlation, grouped data in units of class intervals:

$$\Sigma f(y_C')^2 = \Sigma f_y (d'_{y_c})^2 - \frac{(\Sigma f_y d'_y)^2}{N}.$$

For multiple correlation:

$$\Sigma x_{C1.234}^2 = \Sigma X_{C1\ 234}^2 - \bar{X}_1 \Sigma X_1.$$

For multiple correlation when data are in deviation form:

$$\Sigma x_{C1\ 234}^2 = b_{12\ 34} \Sigma x_1 x_2 + b_{13\ 24} \Sigma x_1 x_3 + b_{14\ 23} \Sigma x_1 x_4.$$

Exponential equation:

$$Y_C = ab^X, \text{ or } \log Y = \log a + X \log b.$$

Compound interest curve is often written $P_n = P_o(1 + r).^n$

$$F = \frac{\bar{\sigma}_1^2}{\bar{\sigma}_2^2} \text{ where } \bar{\sigma}_1^2 \text{ is larger variance.}$$

$$F = \frac{\bar{\sigma}_{y_C}^2}{\bar{\sigma}_{y_s}^2} = \frac{\bar{\sigma}_{C1.234}^2}{\bar{\sigma}_{S1\ 234}^2}.$$

$$F_1\left(\frac{x}{\sigma}\right) = \frac{Ni}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}} : \text{ the normal curve function.}$$

$$F_2\left(\frac{x}{\sigma}\right) = \frac{Ni}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}} \left[\frac{\alpha_3}{2}\left(\frac{x}{\sigma} - \frac{\alpha^3}{3\sigma^3}\right) \right] : \text{ the second term of the Gram-}$$
Charlier series. Cumulative frequencies for Gram-Charlier second approximation curve may be obtained by integration:

$$\int_0^x (fx)\, dx = F_1\left(\frac{x}{\sigma}\right) - \alpha_3 F_2\left(\frac{x}{\sigma}\right).$$

f: number of observations in a class. $\Sigma f = N$.
f_C: a computed frequency.
Factoring, formula for factoring the quadratic equation, $a + bX + cX^2 = 0$:

$$X = \frac{-b \pm \sqrt{b^2 - 4ac}}{2c}.$$

Frequency curve type, criteria of: see α_3; α_4; β_1; β_2; κ_2.
$G = \sqrt[N]{X_1 \cdot X_2 \cdot X_3 \cdots X_N}$: geometric mean, a measure of central tendency.

Geometric mean: see G.

Gompertz curve equation:

$Y_C = ka^{b^x}$, or

$\log Y_C = \log k + b^x \log a.$

$H = \dfrac{N}{\dfrac{1}{X_1} + \dfrac{1}{X_2} + \cdots + \dfrac{1}{X_N}}$: harmonic mean, a measure of central tend-

ency.   Alternate forms of this expression are shown in Chapter IX.

Harmonic mean: see H.

I: irregular movements.

$i = l_2 - l_1$: the class interval.

Index number formulae: see Aggregative index number formulae; Average of relatives index number formulae.

Individual importance of independent variables: see $r_{12\,34}$; $r^2_{12.34}$; $\beta_{12.34}$; $d^2_{12\,34}$; $_{12}r_{34}.$

Infinity: ∞

I.Q.: Intelligence quotient.

K: column.

k: number of samples (used in connection with criterion of likelihood).

$$k = \frac{1}{n}\left[\Sigma_1 Y - \left(\frac{b^n - 1}{b - 1}\right)a\right]$$

$$= \frac{1}{n}\left(\Sigma_1 Y - \frac{\Sigma_2 Y - \Sigma_1 Y}{b^n - 1}\right)$$

$$= \frac{1}{n}\left[\frac{\Sigma_1 Y \Sigma_3 Y - (\Sigma_2 Y)^2}{\Sigma_1 Y + \Sigma_3 Y - 2\Sigma_2 Y}\right]: \text{ the asymptote of a modified exponential}$$

equation.

k: the asymptote of a Gompertz curve equation.

$$\log k = \frac{1}{n}\left[\Sigma_1 \log Y - \left(\frac{b^n - 1}{b - 1}\right)\log a\right]$$

$$= \frac{1}{n}\left(\Sigma_1 \log Y - \frac{\Sigma_2 \log Y - \Sigma_1 \log Y}{b^n - 1}\right)$$

$$= \frac{1}{n}\left[\frac{\Sigma_1 \log Y \,\Sigma_3 \log Y - (\Sigma_2 \log Y)^2}{\Sigma_1 \log Y + \Sigma_3 \log Y - 2\Sigma_2 \log Y}\right].$$

$k = \dfrac{2y_0 y_1 y_2 - y_1^2(y_0 + y_2)}{y_0 y_2 - y_1^2}$: the asymptote of a Pearl-Reed (logistic) equation.

$k = \dfrac{\sigma_{y_s}}{\sigma_y} = \sqrt{\dfrac{\sigma^2_{y_s}}{\sigma^2_y}} = \sqrt{\dfrac{\Sigma y_s^2}{\Sigma y^2}}$: coefficient of alienation.

$$k^2 = \frac{\sigma_{v_s}^2}{\sigma_y^2} = \frac{\Sigma y_s^2}{\Sigma y^2}: \text{coefficient of non-determination.}\quad [r^2 + k^2 = 1.]$$

$$\kappa_2 = \frac{\beta_1(\beta_2 + 3)^2}{4(4\beta_2 - 3\beta_1)(2\beta_2 - 3\beta_1 - 6)}: \text{a general measure of the departure}$$

from normal of a frequency distribution. For a normal distribution, the value of $\kappa_2 = 0$.

Kurtosis: see π_4; α_4; β_2.

$$L = \frac{\sqrt[k]{\bar\sigma_1^2 \times \bar\sigma_2^2 \times \cdots \times \bar\sigma_k^2}}{\frac{1}{k}(\bar\sigma_1^2 + \bar\sigma_2^2 + \cdots + \bar\sigma_k^2)}: \text{criterion of likelihood, the ratio of the geo-}$$

metric mean of several standard deviations to their arithmetic mean. Where samples vary in size, weighted means should be used. See Chapter XIII.

l: the limit of a class.

l_1 is the lower limit; l_2 is the upper limit.

Lag, distribution of, by weighted moving average placed opposite X_{N+1}. Weight formula:

$$\overline{X}_{N+1} = \frac{X_1 + 2X_2 + 3X_3 + 4X_4 + \cdots + NX_N}{1 + 2 + 3 + 4 + \cdots + N}.$$

Likelihood, criterion of: see L.

Log: logarithm.

Logarithmic normal curve: a normal curve using log X values. See Normal curve, and Chapter XI.

Logistic curve (see also Pearl-Reed curve):

$$\frac{1}{Y_c} = k + ab^x.$$

m: number of constants in an equation, number of columns in classified data, or number of strata in a stratified sample.

m: the exponent of a binomial expression.

Mean: see $\overline{X}$, G, H.

Mean deviation: see AD.

Mean square contingency, coefficient of: see C.

Measure of central tendency: see $\overline{X}$, Med, Mo, G, H.

Med: median, a measure of central tendency.   Med $= Q_2$.

Median: see Med.

$$\text{Mo} = l_1 + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) i, \text{ also}$$

$$= \overline{X} - \sigma \, \text{Sk}_\beta, \text{ also}$$

$$= \overline{X} - 3(\overline{X} - \text{Med}) \text{ [an empirical approximation]: mode, a measure}$$
of central tendency.

Mode: see Mo.

Modified exponential equation: $Y_C = k + ab^x$.

Modified polynomial equation: $Y_C = a + bX^{\frac{1}{2}} + \cdots$.

Moment: see ν, π, μ.

μ: moment around the mean with Sheppard's correction. (The μ's and π's are in units of class intervals.)

$\mu_1 = \pi_1 = 0$.

$\mu_2 = \pi_2 - \frac{1}{12}$.

$\mu_3 = \pi_3$.

$\mu_4 = \pi_4 - \frac{1}{2}\pi_2 + \frac{7}{240}$.

Multiple correlation coefficient: $R_{1.234}$. See also $R^2_{1.234}$.

Multiple determination, coefficient of: see $R^2_{1\ 234}$.

Multiple estimating equation:

$X_{C1\ 234} = a_{1\ 234} + b_{12.34}X_2 + b_{13\ 24}X_3 + b_{14\ 23}X_4$, or

$x_{C1\ 234} = b_{12\ 34}x_2 + b_{13.24}x_3 + b_{14\ 23}x_4$.

See also $a_{1.234}$, $b_{12.34}$, $b_{13.24}$, $b_{14.23}$ (constants in a multiple estimating equation).

N: number of items in a sample. In a frequency distribution $N = \Sigma f$.

N_K: number of items in a column.

N_S: number of items in a stratum of a sample.

N': number of items in k samples (used in connection with criterion of likelihood).

n: current or given year of an index. Used only as a subscript in this manner.

n: number of degrees of freedom.

n: number of observations in a group, when partial totals are used in fitting a modified exponential curve, a Gompertz curve, or a logistic curve of modified exponential form. Also, number of items between selected points in fitting a Pearl-Reed (logistic) curve.

n: number of years or periods in the compound interest expression $P_n = P_o(1 + r)^n$.

Non-determination coefficient: see k^2.

Normal curve:

$$Y_C = \frac{Ni}{\sigma\sqrt{2\pi}}\, e^{\frac{-x^2}{2\sigma^2}}.$$

Normal curve with adjustment for skewness (first two terms of Gram-Charlier series):

$$Y_C = \frac{Ni}{\sigma\sqrt{2\pi}}\, e^{\frac{-x^2}{2\sigma^2}} - \left\{ \frac{Ni}{\sigma\sqrt{2\pi}}\, e^{\frac{-x^2}{2\sigma^2}} \left[\frac{\alpha_3}{2}\left(\frac{x}{\sigma} - \frac{x^3}{3\sigma^3} \right) \right] \right\}.$$

ν: a moment about an assumed mean.

$\nu_1 = \dfrac{\Sigma f d'}{N}$: first moment about an assumed mean.

$\nu_2 = \dfrac{\Sigma f(d')^2}{N}$: second moment about an assumed mean.

$\nu_3 = \dfrac{\Sigma f(d')^3}{N}$: third moment about an assumed mean.

$\nu_4 = \dfrac{\Sigma f(d')^4}{N}$: fourth moment about an assumed mean.

o: base year of an index. Used only as a subscript in this manner.

Original data: $T \times C \times S \times I$.

Orthogonal polynomial equation:

$$Y_c = A + BX_1 + CX_2 + DX_3 + \cdots.$$

$$X_{(r+1)} = X_1 X_r - \frac{r^2(N^2 - r^2)}{4(4r^2 - 1)} X_{(r-1)}.$$

Coefficient of $X_r = \dfrac{(2r)!\,(2r+1)!}{(r!)^4 N(N^2 - 1)\cdots(N^2 - r^2)} \Sigma X_r Y.$

N is the number of years or months and r is the degree of the polynomial.

P: a percentile. There are 99 percentiles, $P_1 \cdots P_{99}$.

P: population (when used as a subscript). Also, the number of items in the population.

$P_n = P_o(1 + r)^n$: population at end of period.

P_o: population at beginning of period.

P: price index number. See Aggregative index number formulae; Average of relatives index number formulae.

P_n: price index number for a given year. (P_{36}: price index number for 1936, and similarly for other years.)

P: probability of obtaining a deviation of a given magnitude, or the ratio of the area of the tail (or tails) of a frequency distribution to the entire area under consideration.

P_S: number of items in a stratum of the population.

p: a percentage or a proportion.

$$p_{1+2} = \frac{N_1 p_1 + N_2 p_2}{N_1 + N_2}$$: the estimate of the percentage in the population

made by averaging the percentages in the two samples.

p: the probability of obtaining a success.

$$p = \frac{\overline{X}}{m}$$ for a binomial expression, where m is the exponent, or the number of possible happenings minus 1.

p: price of a commodity.

p_n: price of a commodity in given year.

p_o: price of a commodity in base year.

Parabolic equation:

$$Y_C = aX^b, \text{ or } \log Y = \log a + b \log X.$$

Part correlation coefficient: see $_{12}r_{34}$.

Partial correlation coefficient: $r_{12\ 34}$.   See also $r_{12\ 34}^2$.

Partial determination, coefficient of: see $r_{12\ 34}^2$.

$PE = .6745\sigma$: probable error.

$PE_{\overline{X}} = .6745\ \sigma_{\overline{X}}$: probable error of the arithmetic mean.

Pearl-Reed curve: a type of logistic curve (see also Logistic curve).

$$Y_C = \frac{k}{1 + e^{\,a\,+\,bX}} \text{ (symmetrical curve).}$$

$$Y_C = \frac{k}{1 + e^{\,a\,+\,bX\,+\,cX^2}} \text{ (asymmetrical curve).}$$

Periodic curve: see Sine cosine curve.

π: 3.14159 (circumference of a circle is 2π times the radius).

π: a moment about the mean.

$$\pi_1 = \frac{\Sigma x}{N} = 0: \text{ first moment about the mean.}$$

$$\pi_2 = \sigma^2 = \frac{\Sigma x^2}{N} = \nu_2 - \nu_1^2: \text{ second moment about the mean; variance;}$$
a measure of absolute dispersion.

$$\pi_3 = \frac{\Sigma x^3}{N} = \nu_3 - 3\nu_1\nu_2 + 2\nu_1^3: \text{ third moment about the mean, a meas-}$$
ure of absolute skewness.

$$\pi_4 = \frac{\Sigma x^4}{N} = \nu_4 - 4\nu_1\nu_3 + 6\nu_1^2\nu_2 - 3\nu_1^4: \text{ the fourth moment around the}$$
mean, a measure of absolute kurtosis.

Polynomial equation (see also $a, b, c, d, e \cdot \cdot \cdot$, constants in a polynomial equation):

$$Y_C = a + bX + cX^2 + dX^3 + eX^4 + \cdot \cdot \cdot .$$

Price index number formulae: see Aggregative index number formulae; Average of relatives index number formulae; Purchasing power index number formula.

Price relative: $\dfrac{p_n}{p_o}$.

Purchasing power index number formula:

$$\text{Purchasing power} = \frac{\Sigma \left[(p_o q_o) \dfrac{u_n}{u_o} \right]}{\Sigma p_o q_o}, \text{ where } u = \frac{1}{p}. \text{ (Reciprocal of har-}$$
monic mean of price relatives weighted by base year values.)

Q: quantity index number. See Aggregative index number formulae; Average of relatives index number formulae.

Q_n: quantity index number for a given year. (Q_{36}: quantity index number in 1936, and similarly for other years.)

Q: a quartile. There are three quartiles: Q_1, Q_2, Q_3. Q_2 is the median.

$Q = \dfrac{Q_3 - Q_1}{2}$: quartile deviation or semi-interquartile range, a measure of absolute dispersion.

$q = 1 - p$: a percentage or proportion; also, the probability of obtaining a failure.

$q' = 1 - p'$. See also p'.

q: the quantity of a commodity (used in connection with index numbers).

q_c: the quantity common to two or several periods.

q_n: the quantity of a commodity in the given year.

q_o: the quantity of a commodity in base year.

$q_o + q_n$: the total quantity in two years.

$q_{o,n} = \dfrac{q_o + q_n}{2}$: the average quantity in two years.

q_{o-n}: the average quantity in several years.

Quantity index number formulae: see Aggregative index number formulae; Average of relatives index number formulae.

Quantity relative: $\dfrac{q_n}{q_o}$.

Quartile deviation: see Q.

$R_{1.234}$: correlation of X_1 with $X_{C1.234}$, coefficient of multiple correlation. See also $R^2_{1.234}$.

$R^2_{1.23} = \dfrac{r^2_{12} + r^2_{13} - 2(r_{12})(r_{13})(r_{23})}{1 - r^2_{23}}$: a coefficient of multiple determination.

$$R^2_{1\ 234} = \dfrac{\Sigma x^2_{C1\ 234}}{\Sigma x^2_1} = \dfrac{\Sigma X^2_{C1\ 234} - \overline{X}_1 \Sigma X_1}{\Sigma X^2_1 - \overline{X}_1 \Sigma X_1}$$

$$= 1 - \dfrac{\sigma^2_{S12\ 34}}{\sigma^2_1} = 1 - \dfrac{\Sigma x^2_{S1\ 234}}{\Sigma x^2_1}$$

$$= 1 - [(1 - r^2_{14})(1 - r^2_{13\ 4})(1 - r^2_{12.34})]:$$ a coefficient of multiple determination.

$\bar{R}^2_{1.234} = \dfrac{R^2_{1\ 234}(N - 1) - (m - 1)}{N - m}$: population estimate of $R^2_{1.234}$.

r = rate of change in expression $P_n = P_o(1 + r)^n$.

r: coefficient of correlation, a measure of closeness of relationship between two variables. See also r^2.

For ungrouped data:

$$r = \frac{\sigma_{y_c}}{\sigma_y} = \sqrt{\frac{\sigma_{y_c}^2}{\sigma_y^2}} = \sqrt{1 - \frac{\sigma_{y_s}^2}{\sigma_y^2}} = \sqrt{\frac{\Sigma y_c^2}{\Sigma y^2}} = \sqrt{1 - \frac{\Sigma y_s^2}{\Sigma y^2}}$$

$$= b_{yx} \div \frac{\sigma_y}{\sigma_x} = \frac{\Sigma xy}{\Sigma x^2} \cdot \frac{\sigma_x}{\sigma_y} = \frac{\Sigma\left(\dfrac{x}{\sigma_x} \cdot \dfrac{y}{\sigma_y}\right)}{\Sigma\left(\dfrac{x}{\sigma_x}\right)^2}$$

$$= \frac{1}{N}\Sigma\left(\frac{x}{\sigma_x} \cdot \frac{y}{\sigma_y}\right) = \frac{\Sigma xy}{N\sigma_x\sigma_y} = \sqrt{\frac{(\Sigma xy)^2}{(\Sigma x^2)(\Sigma y^2)}} = \sqrt{b_{yx} \cdot b_{xy}}$$

$$= \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}}.$$

For grouped data, this last expression becomes

$$r = \frac{N\Sigma f d_x' d_y' - (\Sigma f_x d_x')(\Sigma f_y d_y')}{\sqrt{[N\Sigma f_x (d_x')^2 - (\Sigma f_x d_x')^2][N\Sigma f_y (d_y')^2 - (\Sigma f_y d_y')^2]}}$$

$$= \frac{\dfrac{\Sigma f d_x' d_y'}{N} - \dfrac{\Sigma f_x d_x'}{N}\dfrac{\Sigma f_y d_y'}{N}}{\sqrt{\dfrac{\Sigma f_x(d_x')^2}{N} - \left(\dfrac{\Sigma f_x d_x'}{N}\right)^2}\sqrt{\dfrac{\Sigma f_y(d_y')^2}{N} - \left(\dfrac{\Sigma f_y d_y'}{N}\right)^2}}: \text{ useful when esti-}$$

mating equation is to be obtained by use of $b = r\dfrac{\sigma_y}{\sigma_x}$, since de-

nominator gives σ's (in class intervals).

For other grouped data formulae, see r^2.

r^2: coefficient of determination, a measure of closeness of relationship be-
tween two variables. See also r.

For ungrouped data:

$$r^2 = \frac{\sigma_{y_c}^2}{\sigma_y^2} = \frac{\Sigma y_c^2}{\Sigma y^2} = \frac{\Sigma Y_C^2 - \overline{Y}\Sigma Y}{\Sigma Y^2 - \overline{Y}\Sigma Y} = \frac{(a\Sigma Y + b\Sigma XY) - \dfrac{(\Sigma Y)^2}{N}}{\Sigma Y^2 - \dfrac{(\Sigma Y)^2}{N}}$$

$$= 1 - \frac{\sigma_{y_s}^2}{\sigma_y^2} = 1 - \frac{\Sigma y_s^2}{\Sigma y^2} = 1 - \frac{\Sigma Y^2 - \Sigma Y_C^2}{\Sigma Y^2 - \overline{Y}\Sigma Y}.$$

For grouped data:

$$r^2 = \frac{\Sigma(y_c')^2}{\Sigma(y')^2} = \frac{\Sigma f_y (d_{y_c}')^2 - (\Sigma f_y d_y')^2 \div N}{\Sigma f_y (d_y')^2 - (\Sigma f_y d_y')^2 \div N}$$

$$= \frac{N\Sigma f_y (d_{y_c}')^2 - (\Sigma f_y d_y')^2}{N\Sigma f_y (d_y')^2 - (\Sigma f_y d_y')^2}.$$

For further explanation of symbols, see Variance; Total variation; Ex-
plained variation; Unexplained variation; Explained sum of squares.

$$\bar{r}^2 = 1 - \frac{\bar{\sigma}_{v_s}^2}{\bar{\sigma}_y^2} = \frac{r^2(N-1)-1}{N-2} = \frac{r^2(N-1)-(m-1)}{N-m} : \text{estimate of pop-}$$

ulation value of r^2.

$$r_{12.3} = \frac{r_{12} - (r_{13})(r_{23})}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}} : \text{a coefficient of partial correlation. See also}$$

$r_{12\ 34}$ and $r_{12\ 34}^2$.

$r_{12\ 34}$: correlation of $x_{S2.34}$ with $x_{S1.34}$. Coefficient of partial correlation. See also $r_{12\ 34}^2$.

$$r_{12.34} = \sqrt{\frac{\Sigma x_{C1\ 234}^2 - \Sigma x_{C1\ 34}^2}{\Sigma x_{S1\ 34}^2}}$$

$$= b_{12.34} \div \frac{\sigma_{S1\ 234}}{\sigma_{S2\ 134}}, \text{ or } b_{12\ 34} \times \frac{\sigma_{S2\ 134}}{\sigma_{S1\ 234}}$$

$$= \sqrt{b_{12\ 34} \cdot b_{21.34}}$$

$$= \frac{r_{12.3} - (r_{14\ 3})(r_{24\ 3})}{\sqrt{1-r_{14\ 3}^2}\sqrt{1-r_{24\ 3}^2}}, \text{ or } \frac{r_{12\ 4} - (r_{13\ 4})(r_{23.4})}{\sqrt{1-r_{13\ 4}^2}\sqrt{1-r_{23\ 4}^2}}.$$

$$r_{12\ 345\ \cdots\ m} = \frac{r_{12\ 345\ \cdots\ (m-1)} - [r_{1m\ 345\ \cdots\ (m-1)}][r_{2m\ 345\ \cdots\ (m-1)}]}{\sqrt{1-r_{1m\ 345\ \cdots\ (m-1)}^2}\ \sqrt{1-r_{2m\ 345\ \cdots\ (m-1)}^2}} :$$

general formula for the coefficient of partial correlation.

$$r_{12\ 34}^2 = \frac{\Sigma x_{C1\ 234}^2 - \Sigma x_{C1\ 34}^2}{\Sigma x_{S1\ 34}^2} = \frac{\Sigma x_{C1\ 234}^2 - \Sigma x_{C1\ 34}^2}{\Sigma x_1^2 - \Sigma x_{C1\ 34}^2}$$

$$= \frac{\Sigma X_{C1\ 234}^2 - \Sigma X_{C1.34}^2}{\Sigma X_1^2 - \Sigma X_{C1.34}^2} : \text{coefficient of partial determination. See}$$

also $r_{12\ 34}$.

$$1 - r_{12.34}^2 = \frac{1 - R_{1\ 234}^2}{1 - R_{1.34}^2}.$$

$$\bar{r}_{12.34}^2 = \frac{r_{12\ 34}^2(N-m+1)-1}{N-m} : \text{population estimate of } r_{12.34}^2.$$

$_{12}r_{34}$: correlation of X_2 with $(X_1 - b_{13\ 24}X_3 - b_{14\ 23}X_4)$, coefficient of part correlation.

Range of a sine cosine curve: $2\sqrt{A^2 + B^2}$.

Ranked data, correlation of: see ρ, Spearman's formula.

$$\rho = 1 - \frac{6\Sigma D^2}{N(N^2-1)} : \text{Spearman's formula for correlation of ranked data.}$$

ρ (index of non-linear correlation): see ρ^2.

ρ^2: index of determination. A measure of non-linear correlation. Formulae are the same as those for r^2, which are based upon the explanation of variance or variation. Estimating equation may have more con-

stants, or may be for logarithms or reciprocals of X or Y, or may be any combination of these conditions. If logarithms of X and Y are used, the symbols are written $\rho_{\log Y \log X}$, and similarly for other transformations.

$$\bar{\rho}^2 = 1 - \frac{\bar{\sigma}^2_{y_s}}{\bar{\sigma}^2_y} = \frac{\rho^2(N-1) - (m-1)}{N-m}: \text{ estimated population value of } \rho^2.$$

S: deviation from a line of estimation. Used only as a subscript in this sense. Also used as a subscript to refer to a particular stratum in a stratified sample.

S: seasonal movement.

Scatter ratio: anti-log of $\sigma_{\log y_s}$.

Seasonally adjusted data: see Deseasonalized data.

Semi-interquartile range: see Q.

Separate determination coefficient: see $d^2_{12\ 34}$.

Sheppard's method of unlike signs $= \cos U\ 1.8°$, where U is percentage of cases of unlike sign: a measure of correlation of qualitatively classified data.

Σ: summation sign.

ΣX; ΣY: sum of all the X or Y values.

$\displaystyle\sum_{1}^{N_K}$: summation of items 1 through N_K.

$\displaystyle\sum_{1}^{m}$: summation of columns 1 through m.

$$\sum_{1}^{m}\left(\sum_{1}^{N_K} Y\right) = \Sigma Y.$$

Σ_1, Σ_2, Σ_3: partial totals.

ΣY^2: sum of squares of Y values.

ΣY^2_C: see Explained sum of squares.

Σy^2: see Total variation.

Σy^2_C: see Explained variation.

Σy^2_S: see Unexplained variation.

σ: standard deviation, a measure of absolute dispersion.

For ungrouped data:

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{\frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2}.$$

For grouped data:

$$\sigma = \sqrt{\frac{\Sigma f x^2}{N}} = i\sqrt{\frac{\Sigma f(d')^2}{N} - \left(\frac{\Sigma f d'}{N}\right)^2}.$$

$$\bar{\sigma} = \sqrt{\frac{\Sigma x^2}{N-1}}: \text{estimated standard deviation in the population. See also } \bar{\sigma}^2.$$

$$\sigma^2 = \frac{\Sigma x^2}{N}: \text{variance, a measure of absolute dispersion.}$$

$$\bar{\sigma}^2 = \frac{\Sigma x^2}{N-1} = \frac{N}{N-1}\sigma^2: \text{estimated variance in the population.}$$

For ungrouped data:

$$\bar{\sigma}^2 = \frac{\Sigma X^2}{N-1} - \frac{(\Sigma X)^2}{N(N-1)} = \frac{\Sigma X^2 - \dfrac{(\Sigma X)^2}{N}}{N-1} = \frac{\Sigma X^2 - \bar{X}\Sigma X}{N-1}.$$

For grouped data:

$$\bar{\sigma}^2 = \imath^2\left[\frac{\Sigma f(d')^2}{N-1} - \frac{(\Sigma fd')^2}{N(N-1)}\right]$$

$$= \imath^2\left[\frac{\Sigma f(d')^2 - \dfrac{(\Sigma fd')^2}{N}}{N-1}\right].$$

$$\bar{\sigma}_{1+2}^2 = \frac{\Sigma x_1^2 + \Sigma x_2^2}{(N_1-1)+(N_2-1)} = \frac{\Sigma x_1^2 + \Sigma x_2^2}{n_1 + n_2}: \text{estimate of population var-}$$
iance made by averaging the variances of the two samples. Same as estimated population variance within columns when there are two columns.

When $\bar{\sigma}^2$ is estimated from several samples,

$$\bar{\sigma}^2 = \frac{N_1\sigma_1^2 + N_2\sigma_2^2 + \cdots + N_k\sigma_k^2}{(N_1 + N_2 + \cdots + N_k) - k}.$$

$\sigma_a = \sqrt{Npq}:$ standard error of number of occurrences.

$\sigma_{C1\ 234}:$ standard deviation of explained or computed values, multiple cor-
relation. See also $\sigma_{C1\ 234}^2$.

$$\sigma_{C1\ 234}^2: = \frac{\Sigma x_{C1\ 234}^2}{N}: \text{explained variance, multiple correlation. See also ex-}$$
plained variation.

$$\bar{\sigma}_{C1.234}^2 = \frac{\Sigma x_{C1\ 234}^2}{m-1}: \text{explained variance based on degrees of freedom, mul-}$$
tiple correlation.

$$\sigma_\eta = \frac{1-\eta^2}{\sqrt{N-m}}: \text{standard error of the correlation ratio. A rough approxi-}$$
mation.

$$\sigma_{\eta^2-r^2} = 2\sqrt{\frac{\eta^2-r^2}{N}}\sqrt{(1-\eta^2)^2 - (1-r^2)^2 + 1}: \text{standard error of}$$
$\eta^2 - r^2$. An approximation.

$\sigma_{\log}$: standard deviation of the logarithms of a series. May be computed by use of the expression $\sigma_{\log} = .7413 \,(\log Q_3 - \log Q_1)$ for fitting a logarithmic normal curve.

$\sigma_{\text{Med}} = 1.2533\, \sigma_{\bar{x}}$: standard error of the median.

σ_P: standard deviation in the population.

σ_P^2: variance in the population.

$$\sigma_p = \sqrt{\frac{pq}{N}} = \sqrt{\frac{p - p^2}{N}}: \text{ standard error of a percentage.}$$

$$\sigma_{p_1 - p_2} = \sqrt{\sigma_{P_1}^2 + \sigma_{P_2}^2}: \text{ standard error of the difference between two}$$
percentages.

$$\sigma'_{p_1 - p_2} = \sqrt{\frac{p_{1+2}\, q_{1+2}}{N_1} + \frac{p_{1+2}\, q_{1+2}}{N_2}} = \sqrt{p_{1+2}\, q_{1+2}\, \frac{N_1 + N_2}{N_1 N_2}}:$$

the standard error of the difference between two percentages when one estimate of the percentage in the population is made from the two samples.

σ_r: standard error of the coefficient of correlation.

$$\sigma_r = \frac{1 - r_P^2}{\sqrt{N - 1}}, \text{ a rough measure unless } r_P \text{ is small and } N \text{ is large.}$$

$$\sigma_r = \frac{1}{\sqrt{N - 1}}, \text{ when hypothesis is } r_P = 0.$$

$$\sigma_r = \frac{1 - r^2}{\sqrt{N - 2}}, \text{ a rough measure of the sampling error of } r.$$

$$\sigma_{R_{1\,234}} = \frac{1 - R_{1\,234}^2}{\sqrt{N - m}}: \text{ standard error of coefficient of multiple correlation}$$

A rough approximation.

$\sigma_{r_{12\,34}}$: standard error of coefficient of partial correlation.

$$\sigma_{r_{12\,34}} = \frac{1 - r_{P12\,34}^2}{\sqrt{N - m + 1}}, \text{ a rough measure of the sampling error of } r_{12\,34}$$
unless $r_{P12\,34}$ is small and N is large.

$$\sigma_{r_{12\,34}} = \frac{1}{\sqrt{N - m + 1}}, \text{ when hypothesis is } r_{P12\,34} = 0.$$

$$\sigma_{r_{12\,34}} = \frac{1 - r_{12\,34}^2}{\sqrt{N - m}}, \text{ a rough measure of the sampling error of } r_{12.34}.$$

$$\sigma_\rho = \frac{1 - \rho^2}{\sqrt{N - m}}: \text{ standard error of the index of correlation. A rough ap-}$$
proximation.

$\sigma_{S1\ 234}$: standard error of estimate, multiple correlation. See also $\sigma_{\bar{S}1\ 234}^2$.

$$\sigma^2{}_{S1.234} = \frac{\Sigma x_{S1\ 234}^2}{N} = \frac{\Sigma x_1^2 - \Sigma x_{C1\ 234}^2}{N} = \frac{\Sigma X_1^2 - \Sigma X_{C1\ 234}^2}{N}$$

$$= \frac{\Sigma x_1^2\ (1 - r_{14}^2)(1 - r_{13\ 4}^2)\ (1 - r_{12\ 34}^2)}{N}$$

$$= \sigma_1^2\ (1 - r_{14}^2)(1 - r_{13.4}^2)\ (1 - r_{12\ 34}^2): \text{unexplained variance, mul-}$$
tiple correlation.

$$\bar{\sigma}_{S1\ 234}^2 = \frac{\Sigma x_{S1\ 234}^2}{N - m}: \text{estimate of unexplained variance in the population,}$$
multiple correlation.

$$\sigma_\sigma = \frac{\bar{\sigma}}{\sqrt{2N}} = .7071068\ \sigma_{\bar{x}}: \text{standard error of the standard deviation.}$$

If kurtosis is present $\sigma_\sigma = \dfrac{\bar{\sigma}}{\sqrt{2N}}\sqrt{1 - \dfrac{\beta_2 - 3}{2}}$, where β_2 is the value in the population.

$\sigma_{\sigma_1 - \sigma_2} = \sqrt{\sigma_{\sigma_1}^2 + \sigma_{\sigma_2}^2}$: standard error of the difference between two standard deviations (useful when N_1 and N_2 are large).

$$\bar{\sigma}_{\text{of strata means}}^2 = \frac{\overset{m}{\underset{1}{\Sigma}} N_S\ (\bar{X}_S - \bar{X})^2}{N - 1}: \text{estimated population variance of strata}$$
means.

$$\sigma_{\bar{X}\ \text{of a stratified sample}}^2 = \frac{\bar{\sigma}^2}{N} - \frac{\overset{m}{\underset{1}{\Sigma}} N_S\ (\bar{X}_S - \bar{X})^2}{N\ (N - 1)}: \text{the sampling variance of the}$$
mean of a stratified sample.

$\sigma_V = \dfrac{V}{\sqrt{2N}}\sqrt{1 + 2\left(\dfrac{V}{100}\right)^2}$: standard error of coefficient of variation.

$\sigma_{V_1 - V_2} = \sqrt{\sigma_{V_1}^2 + \sigma_{V_2}^2}$: standard error of the difference between two co-efficients of variation.

$\sigma_x^2,\ \sigma_{x_C}^2\ \sigma_{x_S}^2$: similar to corresponding expressions for y.

$$\sigma_{x_1}^2 = \frac{\Sigma x_1^2}{N} = \frac{\Sigma X_1^2 - \bar{X}_1 \Sigma X_1}{N}: \text{variance of the dependent variable in mul-}$$
tiple correlation.

$\sigma_{\bar{X}} = \dfrac{\sigma_P}{\sqrt{N}}$, or approximately $\dfrac{\bar{\sigma}}{\sqrt{N}} = \dfrac{\sigma}{\sqrt{N - 1}}$: standard error of the mean.

$\sigma_{\bar{X}_D}$: standard error of the mean of differences between paired items.

$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\sigma_{\bar{X}_1}^2 + \sigma_{\bar{X}_2}^2}$: standard error of the difference between two means.

$$\sigma'_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\bar{\sigma}^2_{1+2}}{N_1} + \frac{\bar{\sigma}^2_{1+2}}{N_2}}$$

$$= \sqrt{\frac{(N_1 + N_2)(\Sigma x_1^2 + \Sigma x_2^2)}{N_1 N_2[(N_1 - 1) + (N_2 - 1)]}} = \sqrt{\frac{(N_1 + N_2)(\Sigma x_1^2 + \Sigma x_2^2)}{N_1 N_2 (n_1 + n_2)}}:$$

standard error of the difference between two means when one estimate of population variance is made from two samples. This expression is the same as that for $\sigma_{\bar{X}_1 - \bar{X}_2}$ when $N_1 = N_2$.

σ_y^2: variance, or total variance, a measure of absolute dispersion. See also Total variation $= \Sigma y^2$.

For ungrouped data:

$$\sigma_y^2 = \frac{\Sigma y^2}{N} = \frac{\Sigma(Y - \bar{Y})^2}{N} = \frac{\Sigma Y^2}{N} - \left(\frac{\Sigma Y}{N}\right)^2.$$

For grouped data:

$$\sigma_y^2 = i^2\left[\frac{\Sigma f_y (d'_y)^2}{N} - \left(\frac{\Sigma f_y d'_y}{N}\right)^2\right].$$

$$\sigma_y^2 = \sigma_{y_c}^2 + \sigma_{y_s}^2.$$

$\bar{\sigma}_y^2 = \dfrac{\Sigma y^2}{N - 1}$: estimate of total variance in the population of Y, the dependent variable. When only one variable is under consideration, σ usually has no subscript and the deviations are indicated by x. See $\bar{\sigma}$.

$\sigma_{y_c}^2$: explained variance. See also Σy_c^2, or explained variation, and ΣY_c^2, or explained sum of squares.

For ungrouped data:

$$\sigma_{y_c}^2 = \frac{\Sigma y_c^2}{N} = \frac{\Sigma(Y_c - \bar{Y})^2}{N}.$$

For grouped data:

$$\sigma_{y_c}^2 = i^2\left[\frac{\Sigma f(y_c')^2}{N}\right] = i^2\left[\frac{\Sigma f_y (d'_{y_c})^2}{N} - \left(\frac{\Sigma f_y d'_y}{N}\right)^2\right].$$

$\bar{\sigma}_{y_c}^2 = \dfrac{\Sigma y_c^2}{m - 1}$: explained variance based on degrees of freedom.

σ_{y_s}: standard error of estimate. See also $\sigma_{y_s}^2$.

For ungrouped data:

$$\sigma_{y_s} = \sqrt{\frac{\Sigma y_s^2}{N}} = \sqrt{\frac{\Sigma(Y - Y_c)^2}{N}} = \sqrt{\frac{\Sigma Y^2 - \Sigma Y_c^2}{N}} = \sqrt{\frac{\Sigma Y^2 - (a\Sigma Y + b\Sigma XY)}{N}}$$

For grouped data:

$$\sigma_{y_s} = \imath\sqrt{\frac{\Sigma f(y')^2 - \Sigma f(y_c')^2}{N}}$$

$$= i\sqrt{\frac{\Sigma f_y\,(d_y')^2 - \Sigma fy\,(d_{y_c}')^2}{N}}.$$

$$\sigma_{y_s} = \sigma_y\sqrt{1 - r^2}.$$

$\sigma_{y_s}^2$: unexplained variance. See also Unexplained variation $= \Sigma y_s^2$.

For ungrouped data:

$$\sigma_{y_s}^2 = \frac{\Sigma y_s^2}{N} = \frac{\Sigma(Y - Y_c)^2}{N} = \frac{\Sigma Y^2 - \Sigma Y_c^2}{N} = \frac{\Sigma Y^2 - (a\Sigma Y + b\Sigma XY)}{N}.$$

For grouped data:

$$\sigma_{y_s}^2 = i^2\left[\frac{\Sigma f(y')^2 - \Sigma f(y_C')^2}{N}\right]$$

$$= i^2\left[\frac{\Sigma f_y(d_y')^2 - \Sigma f_y\,(d_{y_c}')^2}{N}\right].$$

$\bar{\sigma}_{y_s}^2 = \dfrac{\Sigma y_s^2}{N - m}$: estimate of unexplained variance in the population.

$\sigma_Z = \dfrac{1}{\sqrt{N - m - 1}}$: standard error of Z.

Sine cosine curve equation: $Y_C = \bar{Y} - A\sin\left(\dfrac{360}{T}X\right)^\circ + B\cos\left(\dfrac{360}{T}X\right)^\circ$.

Skewed curve: see Binomial; Logarithmic normal curve; Normal curve with adjustment for skewness.

$\mathrm{Sk} = \dfrac{\bar{X} - \mathrm{Mo}}{\sigma}$, or roughly $\dfrac{3(\bar{X} - \mathrm{Med})}{\sigma}$: a measure of relative skewness.

$\mathrm{Sk}_\beta = \dfrac{\sqrt{\beta_1}\,(\beta_2 + 3)}{2(5\beta_2 - 6\beta_1 - 9)}$: a measure of relative skewness.

$\mathrm{Sk}_{\log} = \dfrac{\log Q_1 + \log Q_3 - 2\log Q_2}{\log Q_3 - \log Q_1}$: a logarithmic measure of relative skewness.

$\mathrm{Sk}_P = \dfrac{P_{10} + P_{90} - 2P_{50}}{P_{90} - P_{10}}$: a measure of relative skewness based upon the percentiles.

$\mathrm{Sk}_Q = \dfrac{Q_1 + Q_3 - 2Q_2}{(Q_3 - Q_1)}$: a measure of relative skewness based upon the quartiles.

Skewness:

For absolute skewness, see π_3.

For relative skewness, see $\alpha_3 = \sqrt{\beta_1}$, β_1, Sk, Sk_β, $\text{Sk}_{\log}$, Sk_P, Sk_Q.

Spearman's formula for correlation of ranked data: see ρ.

Standard deviation: see σ and $\bar{\sigma}$.

Standard error of estimate: see σ_{y_s}, $\bar{\sigma}_{y_s}$, $\sigma_{S1\ 234}$, and $\bar{\sigma}_{S1\ 234}$.

Standard error of a statistical measure: see σ_η, $\sigma_{\eta^2 - r^2}$, σ_{Med}, σ_p, $\sigma_{p_1 - p_2}$, σ_r

σ_ρ, $\sigma_{R_{1.234}}$, $\sigma_{r_{12\ 34}}$, σ_σ, $\sigma_{\sigma_1 - \sigma_2}$, σ_V, $\sigma_{V_1 - V_2}$, $\sigma_{\bar{X}}$, $\sigma_{\bar{X}_1 - \bar{X}_2}$, σ_Z.

Standard score: $\dfrac{X - \bar{X}}{\sigma}$.

Straight line equation: $Y_C = a + bX$.

T: periodicity of a time series in X units.

T: secular trend.

t: ratio of a statistical measure which is distributed normally around a mean of zero to an estimate of the standard error of that measure based on the number of degrees of freedom present.

$$t = \frac{\bar{X} - \bar{X}_P}{\bar{\sigma} \div \sqrt{N}}, \text{ or } \frac{\bar{X} - \bar{X}_P}{\sigma \div \sqrt{N - 1}}$$

$$= \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}_1 - \bar{X}_2}}, \text{ or } \frac{\bar{X}_1 - \bar{X}_2}{\sigma'_{\bar{X}_1 - \bar{X}_2}}$$

$$= r \div \frac{\sqrt{1 - r^2}}{\sqrt{N - m}} = \frac{r\sqrt{N - 2}}{\sqrt{1 - r^2}}$$

$$= \frac{r_{12\ 34} \sqrt{N - m}}{\sqrt{1 - r_{12\ 34}^2}}.$$

Tchebycheff's inequality: $P < \dfrac{1}{\left(\dfrac{x}{\sigma}\right)^2}$, where P is the proportion of items beyond any distance x on both sides of the mean. Applies to any series of data.

Total variance: see σ_t^2.

Total variation: a measure of absolute dispersion. $\Sigma y^2 = \Sigma y_C^2 + \Sigma y_S^2$.

For ungrouped data:

$$\Sigma y^2 = \Sigma Y^2 - \frac{(\Sigma Y)^2}{N} = \Sigma Y^2 - \bar{Y}\Sigma Y.$$

For grouped data:

$$\Sigma y^2 = i^2 \left[\Sigma f_y \, (d_y')^2 - \frac{(\Sigma f_y d_y')^2}{N} \right].$$

For grouped data in units of class intervals:

$$\Sigma f(y')^2 = \Sigma f_y\,(d'_y)^2 - \frac{(\Sigma f_y\,d'_y)^2}{N}.$$

For multiple and partial correlation:

$$\Sigma x_1^2 = \Sigma X_1^2 - \overline{X}_1 \Sigma X_1.$$
$$\Sigma x_1^2 = \Sigma x_{C1\ 234}^2 + \Sigma x_{S1\ 234}^2.$$

Unexplained variance: see $\sigma_{y_s}^2$.

Unexplained variation:

For simple correlation, ungrouped data:

$$\Sigma y_S^2 = \Sigma(Y - Y_C)^2 = \Sigma Y^2 - \Sigma Y_C^2,\ \text{or}\ \Sigma y^2 - \Sigma y_C^2.$$

For grouped data:

$$\Sigma f y_S^2 = i^2[\Sigma f(y')^2 - \Sigma f(y'_c)^2] = i^2[\Sigma f_y(d'_y)^2 - \Sigma f_y(d'_{y_c})^2].$$

For grouped data in units of class intervals:

$$\Sigma f(y'_S)^2 = \Sigma f(y')^2 - \Sigma f(y'_C)^2 = \Sigma f_y(d'_y)^2 - \Sigma f_y(d'_{y_c})^2.$$

For multiple and partial correlation:

$\Sigma x_{S1\ 234}^2 = \Sigma x_1^2 - \Sigma x_{C1\ 234}^2 = \Sigma X_1^2 - \Sigma X_{C1.234}^2.$ See also Explained sum of squares: ΣY_C^2; $\Sigma f_y(d'_{y_c})^2$; $\Sigma X_{C1.234}^2$. See also Variation within columns.

$V = \dfrac{\sigma}{\overline{X}}$: coefficient of variation, a measure of relative dispersion.

$V = \dfrac{\Sigma p_n q_n}{\Sigma p_o q_o}$: relative aggregate value.

$v = pq$: value of a commodity.

$v_n = p_n q_n$: value of a commodity in given year.

$v_o = p_o q_o$: value of a commodity in base year.

Variance: second moment about the mean, or square of standard deviation. See σ_y^2, $\sigma_{x_1}^2$ = total variance.

See $\sigma_{y_c}^2$, $\sigma_{C1.234}^2$ = explained variance.

See $\sigma_{y_s}^2$, $\sigma_{S1\ 234}^2$ = unexplained variance.

Variance, population estimate:

Total . .. . $\dfrac{\Sigma Y^2 - \overline{Y}\Sigma Y}{N - 1}$, or $\dfrac{\Sigma Y^2 - [(\Sigma Y)^2 \div N]}{N - 1}$.

Between columns..... $\dfrac{\overset{m}{\underset{1}{\Sigma}}\left(\overline{Y}_K \overset{N_K}{\underset{1}{\Sigma}}Y\right) - \overline{Y}\Sigma Y}{m - 1}$, or

$$\dfrac{\overset{m}{\underset{1}{\Sigma}}\left[\left(\overset{N_K}{\underset{1}{\Sigma}}Y\right)^2 \div N_K\right] - \left[(\Sigma Y)^2 \div N\right]}{m - 1}.$$

Within columns ... . $\dfrac{\Sigma Y^2 - \overset{m}{\underset{1}{\Sigma}}\left(Y_K \overset{N_K}{\underset{1}{\Sigma}} Y\right)}{N - m}$, or

$$\dfrac{\Sigma Y^2 - \left[\overset{m}{\underset{1}{\Sigma}}\left(\overset{N_K}{\underset{1}{\Sigma}} Y\right)^2 \div N_K\right]}{N - m}.$$

Variation, coefficient of: see V.

Variation: sum of squared deviations. See Total variation; Explained variation; Unexplained variation; Variation between columns; Variation within columns.

Variation between columns or groups (explained variation):

Ungrouped as to Y values:

$$\overset{m}{\underset{1}{\Sigma}}\left[N_K\,(\overline{Y}_K - \overline{Y})^2\right] = \overset{m}{\underset{1}{\Sigma}}\left[\dfrac{\left(\overset{N_K}{\underset{1}{\Sigma}} Y\right)^2}{N_K}\right] - \dfrac{(\Sigma Y)^2}{N} = \overset{m}{\underset{1}{\Sigma}}\left(\overline{Y}_K \overset{N_K}{\underset{1}{\Sigma}} Y\right) - \overline{Y}\Sigma Y.$$

For data grouped as to Y values, deviations in units of class intervals:

$$\overset{m}{\underset{1}{\Sigma}}\,N_K\left(\dfrac{\overline{Y}_K - \overline{Y}}{i}\right)^2 = \overset{m}{\underset{1}{\Sigma}}\left[\dfrac{\left(\overset{N_K}{\underset{1}{\Sigma}} f_y\, d'_y\right)^2}{N_K}\right] - \dfrac{(\Sigma f_y\, d'_y)^2}{N}.$$

Variation within columns or groups (unexplained variation):

Ungrouped as to Y values:

$$\overset{m}{\underset{1}{\Sigma}}\left[\overset{N_K}{\underset{1}{\Sigma}}(Y - \overline{Y}_K)^2\right] = \Sigma Y^2 - \overset{m}{\underset{1}{\Sigma}}\left[\dfrac{\left(\overset{N_K}{\underset{1}{\Sigma}} Y\right)^2}{N_K}\right] = \Sigma Y^2 - \overset{m}{\underset{1}{\Sigma}}\left(\overline{Y}_K \overset{N_K}{\underset{1}{\Sigma}} Y\right).$$

(Same as $\bar{\sigma}^2_{1+2}$ when there are two columns.)

For data grouped as to Y values, deviations in units of class intervals:

$$\overset{m}{\underset{1}{\Sigma}}\left[\overset{N_K}{\underset{1}{\Sigma}}\left(\dfrac{Y - \overline{Y}_K}{i}\right)^2\right] = \Sigma f_y(d'_y)^2 - \overset{m}{\underset{1}{\Sigma}}\left[\dfrac{\left(\overset{N_K}{\underset{1}{\Sigma}} f_y\, d'_y\right)^2}{N_K}\right].$$

X: a variable, usually the independent variable; also the mid-value of a class in a frequency distribution.

X_1: the dependent variable when there are more than two variables. Used in multiple and partial correlation.

$X_1, X_2, X_3 \cdots X_N$: different values or observations of variable X. (Occasionally $X_a, X_b, X_c \cdots X_N$ are used for this concept, as in Appendix B, sections XII–1 and XII–2.)

X_1, X_2, X_3, etc.: independent variables in an orthogonal polynomial equation.   For general expression for X_r, see Orthogonal polynomial equation.

$X_1 = X.$

$X_2 = X_1^2 \dfrac{N^2 - 1}{12}.$

$X_3 = X_1^3 \dfrac{3N^2 - 7}{20} X_1.$

X_2, X_3, X_4, etc.: independent variables in multiple estimating equation

$\overline{X} = \dfrac{\Sigma X}{N} = \overline{X}_d + \left(\dfrac{\Sigma f d'}{N}\right)i$: arithmetic mean, a measure of central tendency.   (Also the best estimate of $\overline{X}_P$.)

$\overline{X}_D$: mean of differences between paired items.

$\overline{X}_d$: an assumed mean.

$\overline{X}_{\log}$: mean of logarithms of X values.   May be computed by use of the expression $\overline{X}_{\log} = \dfrac{\log Q_1 + \log Q_3 + 1.2554 \log Q_2}{3.2554}$ for fitting a logarithmic normal curve.

$\overline{X}_P$: population mean.

$\overline{X}_S$: mean of stratum in a stratified sample.

$x = X - \overline{X}$: a deviation from the mean.

$x' = \dfrac{X - \overline{X}}{i}$: a deviation from the mean in units of class intervals.

x_0, x_1, x_2: selected equidistant points, used in fitting Pearl-Reed (logistic) curve.

$\dfrac{x}{\sigma}$: deviation from mean in units of standard deviations.   Used also for deviation of any computed measure from its population value, from a hypothetical value or from another computed sample value, in units of standard errors of that measure.

Y: a variable, usually the dependent variable.

Y_C: a computed Y value.

Y_0: ordinate at the mean.   Maximum ordinate in case of normal curve.

$\overline{Y}_K$: the mean of a particular column.

$y = Y - \overline{Y}.$

$y' = \dfrac{Y - \overline{Y}}{i}.$

y_0, y_1, y_2: selected Y values, used in fitting Pearl-Reed (logistic) curve.

$Z = \frac{1}{2}[\log_e (1 + r) - \log_e (1 - r)] = \frac{1}{2} \log_e \dfrac{1 + r}{1 - r} = 1.15129 \log_{10} \dfrac{1 + r}{1 - r}$: a transformation of r made in order to obtain an approximately normal sampling distribution.

$$z = \log_e \bar{\sigma}_1 - \log_e \bar{\sigma}_2 = \log_e \frac{\bar{\sigma}_1}{\bar{\sigma}_2},$$

$$= 2.30259(\log_{10} \bar{\sigma}_1 - \log_{10} \bar{\sigma}_2) = 2.30259 \log_{10} \frac{\bar{\sigma}_1}{\bar{\sigma}_2},$$

$$= \tfrac{1}{2}(\log_e \bar{\sigma}_1^2 - \log_e \bar{\sigma}_2^2) = \tfrac{1}{2} \log_e \frac{\bar{\sigma}_1^2}{\bar{\sigma}_2^2},$$

$$= 1.15129(\log_{10} \bar{\sigma}_1^2 - \log_{10} \bar{\sigma}_2^2) = 1.15129 \log_{10} \frac{\bar{\sigma}_1^2}{\bar{\sigma}_2^2},$$

$$= 1.15129 \log_{10} \frac{\bar{\sigma}_{y_c}^2}{\bar{\sigma}_{y_s}^2} = 1.15129 \log_{10} \frac{\bar{\sigma}_{C1\ 234}^2}{\bar{\sigma}_{S1\ 234}^2}.$$

INDEX

INDEX